Applications of Missing Feature Theory to Speaker Recognition

by

Michael Thomas Padilla

B.S., Electrical Engineering (1997)
University of California, San Diego

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2000

Author .....
Department of Electrical Engineering and Computer Science
January 31, 2000

Certified by. .....
Thomas F. Quatieri
Senior Staff, MIT Lincoln Laboratory
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

This work was sponsored by the Department of the Air Force. Opinions, interpretations, conclusions, and recommendations are those of the author and not necessarily endorsed by the United States Air Force.

Applications of Missing Feature Theory to Speaker Recognition

by

Michael Thomas Padilla

Submitted to the Department of Electrical Engineering and Computer Science
on January 31, 2000, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

An important problem in speaker recognition is the degradation that occurs when speaker models trained with speech from one type of channel are used to score speech from another type of channel, known as *channel mismatch.* This thesis investigates various channel compensation techniques and approaches from missing feature theory for improving Gaussian mixture model (GMM)-based speaker verification under this mismatch condition. Experiments are performed using a speech corpus consisting of "clean" training speech and "dirty" test speech equal to the clean speech corrupted by additive Gaussian noise. Channel compensation methods studied are cepstral mean subtraction, RASTA, and spectral subtraction. Approaches to missing feature theory include missing feature compensation, which removes corrupted features, and missing feature restoration which predicts such features from neighboring features in both frequency and time. These methods are investigated both individually and in combination. In particular, missing feature compensation combined with spectral subtraction in the discrete Fourier transform domain significantly improves GMM speaker verification accuracy and outperforms all other methods examined in this thesis, reducing the equal error rate by about 10% more than other methods over a SNR range of 5-25 dB. Moreover, this considerably outperforms a state-of-the-art GMM recognizer for the mismatch application that combines missing feature theory with spectral subtraction developed in a mel-filter energy domain. Finally, the concept of missing restoration is explored. A novel linear minimum mean-squared-error missing feature estimator is derived and applied to pure vowels as well as a clean/dirty verification trial. While it does not improve performance in the verification trial, a large SNR improvement for features estimated for the pure vowel case indicate promise in the application of this method.

Thesis Supervisor: Thomas F. Quatieri
Title: Senior Staff, MIT Lincoln Laboratory

Acknowledgments

I would like to foremost thank Tom Quatieri, my advisor and mentor, for allowing me to work with him as a research assistant and for always very graciously offering me his time and effort in guiding me along the way towards the completion of this thesis. It has truly been not only very educational, but also very enjoyable working with Tom. One of the most important things I've learned from Tom is the great value of creativity in research. Thank you very much.

I also sincerely thank Doug Reynolds and Jack McLaughlin for also always being there to offer me their time and invaluable comments and suggestions. I have a feeling that the mutual interest that Jack and I have in traveling may lead to us to meet unexpectedly at some far flung corner of the world someday. A big thank you also goes out to Cliff Weinstein and Marc Zissman for accepting me into the Lincoln Laboratory Group 62 and the US Air Force for offering the funding that made it all possible. I also thank the rest of the Group 62 members for their friendliness and for making the lab a very enjoyable working environment.

On a more personal note I thank all of my dear friends, both in the US and abroad, for all of their encouragement ever since I can remember. The joy their friendship has brought to my life has enhanced the academic side and made it all that much easier.

Last, but certainly not least, I thank my family - my parents Manuel and Martha, my grandmother Laura ("Golden Oldie"), my uncle Steve, and all my other relatives scattered around the US for all of the love and support they have offered me throughout my life.

Contents

1 General Introduction **12**

1.1 Introduction to Speaker Recognition 12

1.2 The GMM Speaker Recognition System 14

1.3 Mel-Cepstrum and Mel-filter Energy Speech Features 19

1.4 The Speaker Mismatch Problem in Speaker Verification 21

1.5 Problem Addressed 22

1.6 Contribution of Thesis 25

1.7 Organization of Thesis 26

2 Channel Compensation and Techniques for Robust Speaker Recognition **28**

2.1 Cepstral Mean Subtraction 30

2.2 RelAtive SpecTrA 32

2.2.1 Introduction and Basic System 32

2.2.2 Lin-Log RASTA 34

2.3 Delta Cepstral Coefficients 36

2.4 Results with TSID Corpus 37

3 Spectral Subtraction **40**

3.1 Mel-Filter Energy Domain Spectral Subtraction 41

3.2 Soft $|DFT|$ Domain Spectral Subtraction 43

3.3 Results for Spectral Subtraction 46

4 Missing Feature Theory **49**

4.1 Missing Feature Theory 50

4.2 Missing Feature Compensation 50

4.3 Missing Feature Detection 56

4.4 Results for Missing Feature Compensation 59

5 Cascade Noise Handling Systems **65**

5.1 RASTA with Spectral Subtraction 66

5.2 RASTA with Missing Feature Compensation 66

5.3 $|DFT|$ Domain Spectral Subtraction with Missing Feature Compensation 69

5.4 RASTA with Spectral Subtraction and Missing Feature Compensation 69

6 Missing Feature Restoration **72**

6.1 Previous Work in MFR 73

6.1.1 Integrated Speech-Background Model 73

6.1.2 Mean Estimation 74

6.1.3 Results 75

6.2 Time-Frequency Linear Minimum Mean-Squared Error Missing Feature Estimation 75

6.3 Proposed System for Missing Feature Restoration 81

6.4 Results Using Linear MMSE Missing Feature Restoration 90

7 Conclusions **94**

7.1 Summary of Thesis 94

7.2 Suggestions for Future Research 97

List of Figures

1-1 Sample DET curve. 14

1-2 Example: Distribution of features from states "A" and "B". 15

1-3 Graphical representation of speaker models. 17

1-4 Triangular Mel-Scale Filterbank. 20

1-5 Representative "clean" (top) and "dirty" (bottom) speech time-domain waveforms from the TSID corpus. Both are from the same utterance. 23

1-6 Representative "clean" (top) and "dirty" (bottom) speech spectrograms from the TSID corpus. Both are from the same utterance. 23

1-7 DET curve showing error rate performance for baseline tests on TSID corpus. 24

2-1 The frequency response of the RASTA filter $H(\omega)$. 33

2-2 The impulse response of the RASTA filter h[n]. 33

2-3 Comparison of various equalization techniques. 38

3-1 Plot showing EER vs. SNR (AWGN) for both the baseline case as well as with linear mel-filter energy domain spectral subtraction. It is seen that the performance relative to the baseline depends on the SNR of the additive noise, gains over baseline occuring at lower SNR levels. 47

3-2 Plot showing EER vs. SNR (AWGN) for both the baseline case as well as with $|DFT|$ domain spectral subtraction. The application of spectral subtraction in this domain is seen to improve performance over the baseline case significantly at all SNR levels, going from about a 4% improvement at high SNRs to a near 10% improvement at low SNRs. For comparison, results for RASTA processing are also shown. 48

4-1 This plot shows the background model normalized log probability scores without MFC versus the corresponding values with MFC. The hope is that normalization of speakers' log probability scores with the background model's score will help to properly normalize the probability space in cases where MFC is applied. In this plot, the scores where MFC was applied resulted from removing the 10^{th} feature from every frame. It is seen that corresponding scores with and without MFC tend to be highly correlated, supporting the claim the background model normalization helps correct for improperly normalized pdf's when MFC is applied. Similar results were seen when slightly more features were removed in a controlled fashion as well. 54

4-2 Illustration of DET performance with 10^{th} linear mel-filter energy feature floored and with the same feature removed via MFC. MFC is seen to recover the baseline performance. DET curves with 20 dB of additive noise are given for reference. 55

4-3 Trade-off between corrupted features and removing speech information in MFC. This curve shows EER performance with and without MFC for an increasing number of randomly corrupted or removed mel-energy features. 57

4-4 Average correlation between the perfect mf detector ($\alpha = 3.0$) and nonperfect mf detectors ($\alpha = 1.0$ and 3.0). Shows that the average correlation for the $\alpha = 3.0$ nonperfect detector is > 0.7 for most of the features, close to 0.9 for the highest ten features. The nonperfect detector with $\alpha = 1.0$ has been shown for reference. 60

4-5 Plot showing the average number a particular feature is declared missing per frame by the perfect mf detector($\alpha = 3.0$) and nonperfect mf detectors ($\alpha = 1.0$ and 3.0). The perfect and $\alpha = 3.0$ nonperefect detectors are seen to have produced almost identical results. Results are for 20 dB AWGN. . 61

4-6 Plot showing the frequency per frame that a given total number of missing features are detected by the perfect mf detector(α = 3.0) and nonperfect mf detectors (α = 1.0 and 3.0). The results for the perfect and α = 3.0 nonperefect detectors are seen to be almost identical. Results are for 20 dB AWGN. 62

4-7 EER vs. SNR for MFC with both perfect and nonperfect applications of MFC as well as the baseline case. Both the nonperfect and perfect missing feature detectors are shown to have nearly identical results at all SNR levels. At lower SNRs MFC is seen to degrade performance slightly (< 1%), while at higher SNR levels above approximately 17 dB it starts to improve performance by close to 3.5%. 64

5-1 EER performance as a function of SNR for cascade systems using RASTA and one of linear mel-filter energy or $|DFT|$ SS systems. For comparison the baseline system, the RASTA system, and the $|DFT|$ system are plotted as well. It is seen that when linear mel-filter energy SS is used with RASTA that performance is worse than straight RASTA at all SNR levels except than the lowest, at 5 dB. This is in contrast to the system using RASTA with $|DFT|$ SS, which is seen to do better than either RASTA or $|DFT|$ SS used alone at almost all SNRs. For this system EER performance is typically 8% better than the baseline system. 67

5-2 Performance with a system combining RASTA with missing feature compensation. Performance curves for the baseline and pure RASTA systems are shown for comparison. It is seen that the RASTA+MFC system underperforms both the baseline and pure RASTA systems, particularly the pure RASTA system. Only at high SNR levels of 20 dB and above are any benefits seen, but then only by about 1%. 68

5-3 Combination systems of $|DFT|$ domain SS + missing feature compensation and linear mel-filter energy SS + missing feature compensation. Also shown are the baseline and pure SS systems' performance for comparison. The $|DFT|$ domain SS + MFC system substantially outperforms the linear mel-filter energy SS + MFC system as well as the other systems shown, at all SNR values. It is seen to perform better than the baseline system by approximately 15%. The linear mel-filter energy SS + MFC system, on the other hand, only achieves performance roughly halfway between these two. 70

5-4 RASTA processing in combination with missing feature compensation and linear mel-filter energy SS or $|DFT|$ domain SS. The system with the $|DFT|$ domain SS is seen to do better than the other system as well as baseline for at all noise levels. The system with linear mel-filter energy SS only improves over the baseline at low SNR values, hurting performance at the higher ones. 71

6-1 3-Dimensional plot of linear mel-filter energy feature trajectories from a clean speech file. 82

6-2 3-Dimensional plot of logarithmic mel-filter energy feature trajectories from a clean speech file. 83

6-3 Example clean linear mel-filter energy time trajectory. It is seen that the waveform is far from random, having the property that features close in time tend to have similar values and a perceivable underlying deterministic nature. 83

6-4 Representative autocorrelation function for several linear mel-filter energy features. Speech is taken from a clean speech file. 84

6-5 Representative crosscorrelation function for several linear mel-filter energy features. Speech is taken from a clean speech file. 84

6-6 Representative autocorrelation function for several linear mel-filter energy features from a pure vowel (/ε/). It is seen that the amount of autocorrelation for each trajectory is rather high, compared to the overall average. . . 86

6-7 Representative crosscorrelation function for several linear mel-filter energy features and their neighbors from a pure vowel (/ε/). It is seen that the amount of crosscorrelation for each trajectory is rather high, compared to the overall average. 86

6-8 Representative autocorrelation function for several linear mel-filter energy features from a pure fricative (/∫/). It is seen that the amount of autocorrelation for each trajectory is rather low, compared to the overall average and the pure vowel. 87

6-9 Representative crosscorrelation function for several linear mel-filter energy features and their neighbors from a pure fricative (/∫/). It is seen that the amount of crosscorrelation for each trajectory is rather low, compared to the overall average and the pure vowel. 87

6-10 EER values for clean/dirty speaker verification task where missing features are detected and replaced, when possible, by using the linear MMSE missing feature estimator. Performance is seen to degrade considerably in comparison to the baseline case. 92

List of Tables

2.1 Results applying various channel compensation techniques to TSID data for train clean/test dirty case. 37

6.1 Results of applying proposed linear MMSE missing feature restoration system to a *pure vowel* /ε/. Columns indicate additive noise level, percentage of speech features declared to be missing out of all speech features, percentage of features restored out of all speech features, the SNR of all features restored prior to restoration, and the SNR of all features restored after restoration. Results are for an artificially corrupted clean vowel recording. 91

Chapter 1

General Introduction

In this chapter the problem of *speaker recognition* is introduced along with various subareas of research and associated terminology. This leads into a description of the currently most commonly used statistical speaker model, the *Gaussian mixture model*, as well as the type of speech representation typically used by this model. Finally, the general problem addressed by this thesis is described.

1.1 Introduction to Speaker Recognition

The goal in speaker recognition is to recognize a person from his or her voice. Within this broad description are two specific problems: *speaker identification* and *speaker verification*. Speaker identification attempts to associate an unknown voice with a particular known voice taken from a known set of voices. Speaker verification, on the other hand, tries to determine if an unknown voice matches a particular known voice. The speech used for these tasks can be either a known phrase, termed *text-dependent*, or can be a completely unconstrained phrase, termed *text-independent*. Typical applications for speaker recognition are reconnaissance, forensics, access control, automated telephone transactions, speech data management, etc.

Both speaker identification and speaker verification tasks can be thought of as consisting of two stages: *training* and *testing*. In training, the goal is to efficiently (in terms of both computational complexity as well as a minimum of information redun-

dancy) extract from the speech signal some values that represent characteristics that are as unique to that speaker as possible and contain very little of the environment's (the "channel") characteristics, since the acoustic environment in which the speech was recorded might otherwise be falsely included into the speaker models. While to humans the identity of the speaker of a given utterance appears through a complex combination of both high-level cues (such as semantics and diction), mid-level cues (such as prosodics), and low-level cues (such as the acoustic nature of the speech waveform including its frequency domain behavior, pitch, etc.), automatic speaker recognition systems have so far been unable to efficiently and effectively take advantage of any information other than low-level acoustic cues since these have been the easiest to automatically (without human supervision) extract and apply. The most commonly employed acoustic cues are spectral features such as formant trajectories, the resonant frequencies of the vocal tract, and the speaker's pitch.

In describing the performance of speaker ID and verification systems there are typically two types of error measures that are focused on. The first, the *miss probability*, represents the case where a true *claimant*, the speaker who is "claiming" to have produced the test utterance, is incorrectly rejected by the system. The second error measure, the *false acceptance probability*, represents the case where an *impostor*, a speaker who falsely claims to have produced the test utterance, is incorrectly verified by the system as having produced the speech. The relative amounts of both types of errors that occur is determined by the threshold setting used in a maximum likelihood ratio test, which is applied in a manner similar to its use in detection problems in digital communications theory. In a manner similar to how receiver operating characteristic curves are used in communications, in speaker verification problems *detection error trade-off (DET)* curves are used in which false acceptance probabilities are plotted versus miss probabilities for a wide range of possible threshold settings, indicating all the different operating points of the system that are achievable. While the overall behavior of DET curves is of interest, the *equal error rate (ERR)* point, where the two types of errors are equal, is often focused on as a good representative measure that does not bias the system towards a particular type of error in comparing

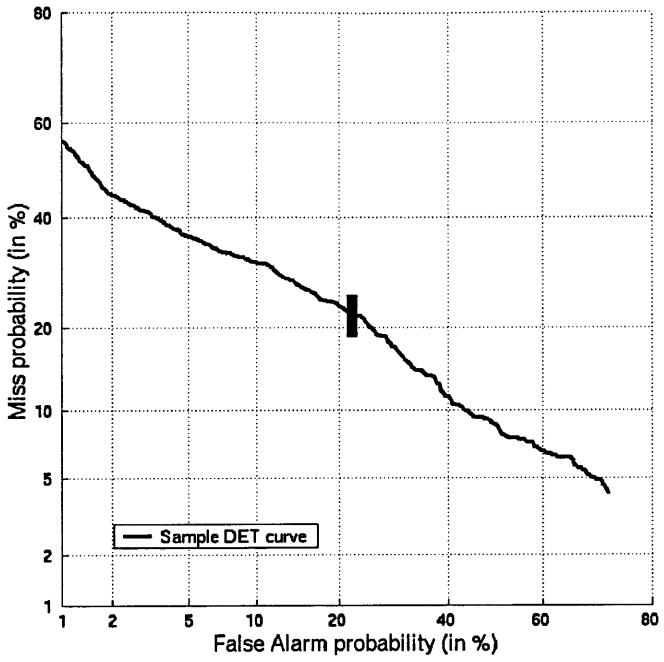

Figure 1-1: Sample DET curve.

different speech systems. A example DET curve is given in figure 1-1.

1.2 The GMM Speaker Recognition System

In most modern speech systems, features are extracted from speech windowed with a 20 ms window and the window is typically advanced at 10 ms intervals. To avoid extracting features of the channel and its noise characteristics, it is essential that a speech detector be used to estimate and indicate which of the windowed frames contain mostly speech and which contain mostly channel noise. For frames that are labeled as speech, the system takes the windowed speech and performs *short-time Fourier transform (STFT)* based analysis on the segment. This is a filterbank analysis which reduces the spectral representation, typically to the commonly employed *mel-cepstral coefficients*, which will be described in the following section. Following this, there will often be a channel equalization stage to mitigate the effect of the channel. Common equalization techniques, which will be described later, are *RASTA* and

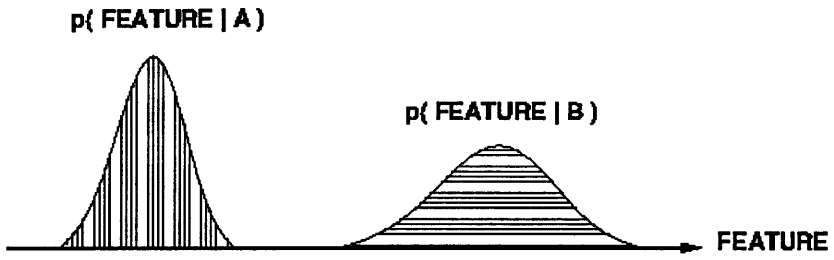

Figure 1-2: Example: Distribution of features from states "A" and "B".

cepstral mean subtraction.

It is of interest how these speech features (typically mel-cepstral features) are associated with different speakers in a quantitative and statistical manner, allowing the construction of mathematical speaker models. In answering this question, speech researchers have depended on much of the work previously done in the area of statistical pattern recognition. Since the primary speaker-dependent acoustic information employed is taken from the spectrum reflecting vocal tract shapes, it is preferred to create speaker models that in some way capture those shapes as manifested in the speech features. To understand what approach is taken, assume that there are two states "A" and "B" (either distinct speakers or vocal tract states of a single speaker) and that each class produces features vectors with a certain probability distribution, as shown in figure 1-2 [1].

If we assume that the underlying distributions of the feature vectors in each class are Gaussian, then it is possible to train the model parameters of the underlying probability density functions (pdf's) in an unsupervised manner not requiring any human supervision by using the expectation-maximization (EM) algorithm [10]. Using the state models, a new feature may be classified as follows

$$p(x|A) \underset{\text{x is from state "B"}}{\overset{\text{x is from state "A"}}{\gtrless}} p(x|B),$$

assuming equally likely classes "A" and "B". It is thus seen that this is essentially a hypothesis testing problem.

[1] Figures 1-2, 1-3, and 1-4 were kindly provided by Doug Reynolds.

In the statistical speaker model, the speaker is regarded as a random source producing the observed speech feature vectors $\mathbf{X}$ and it is the "state" that the speaker's vocal tract is in, as in the example using states A and B above, that determines the general distribution of these vectors. While it is understood that the probability of the speaker being in any one of the states describing his or her speech production mechanism nor that the transition probabilities between the different states is clearly not uniform, for computational ease an assumption of uniform state and state transition probabilities is made. With this assumption, it may be shown that the pdf of the observed speaker is a Gaussian mixture model (GMM)[10]. If an M-state statistical speaker is assumed, the resulting GMM speaker model is given by

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{X})$$

where

$$b_i(\mathbf{X}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{X} - \mu_i)^T \Sigma_i^{-1}(\mathbf{X} - \mu_i)\}$$

In the above expression μ_i and Σ_i are the mean vector and covariance matrix for speaker state i, D is the dimensionality of the feature vectors, and p_i is the probability of being in state i. Note that the set of quantities

$$\lambda = (p_i, \mu_i, \Sigma_i), \quad \text{for i} = 1,...,\text{M}$$

represents the parameters for each state of the speaker model for speaker λ, and hence constitutes a speaker model. This concept is shown graphically in figure 1-3. It is thus observed that in this model the probability of the observed feature vector $\mathbf{X}$ from speaker model λ is the sum of pdf's for each of the hidden and unknown states, appropriately weighted by the probability p_i of the speaker being in that state. Given this resulting speaker model $p(\mathbf{X}|\lambda)$, a quantitative score for the likelihood that an unknown feature vector was generated by a particular speaker may be calculated. The model parameter estimates (p_i, μ_i, Σ_i), i=1,...,M are obtained by the EM algorithm

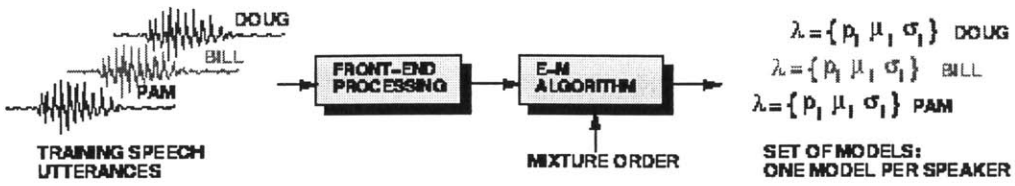

Figure 1-3: Graphical representation of speaker models.

mentioned earlier.

Although the speaker verification task requires only a binary decision of whether the claimant should be accepted or rejected, it is a more involved problem than the simple identification task[10]. This is a result of the fact that the class of "not speaker" is not very clearly defined. In the case of speaker identification there is a known well-defined set of speaker models which the characteristics of the test utterance may be compared against. However, in the speaker verification task there are two competing classifications: "claimant", which has a clearly defined model, and "not claimant", which is very vague and poorly defined since it could be any speaker other than the claimant. Essentially, the speaker verification system must decide if the unknown voice belongs to the claimed speaker, with a well-defined model, or to some other speaker, with an ill-defined model. This latter model is typically referred to as the *background model.* Because it is a binary decision there are two types of errors as mentioned earlier: false rejections (misses) in which case the system rejects the true speaker and false acceptances where an imposter is accepted. A model of the possible imposter speakers must be used to perform the optimum likelihood ratio test that decides between "is claimant" and "is not claimant". For an unknown speech file $\mathbf{X}_T = [\mathbf{X}_1\ \mathbf{X}_2\ ...\ \mathbf{X}_T]$, T representing the number of frames or speech vectors $\mathbf{X}_t$ in the speech file, a claimant with model λ_c, and a model that encompasses all possible nonclaimant speakers λ_{bkgd}, the likelihood ratio is given by

$$\frac{Pr(\mathbf{X}_T \text{ is from claimant})}{Pr(\mathbf{X}_T \text{ is not from claimant})} = \frac{Pr(\lambda_c|\mathbf{X}_T)}{Pr(\lambda_{bkgd}|\mathbf{X}_T)}.$$

With the application of Bayes' Rule and the assumption of equally likely prior probabilities for the claimant and background speakers, the above expression may be simplified with the help of the logarithm to

$$\Lambda(\mathbf{X}_T) = \log[p(\mathbf{X}_T|\lambda_c)] - \log[p(\mathbf{X}_T|\lambda_{bkgd})].$$

The likelihood ratio is compared with a threshold θ and the claimant is accepted if $\Lambda(\mathbf{X}_T) > \theta$ and rejected if $\Lambda(\mathbf{X}_T) < \theta$. The likelihood ratio measures the degree by which the claimant model is statistically closer to the observed speech $\mathbf{X}_T$ than the background model. The decision threshold θ is determined by the desired trade-off between the false acceptance and false rejection rates. The components of the likelihood ratio test are computed as follows: for the likelihood that the speech came from the claimant, the likelihood is calculated from the frames from $\mathbf{X}_T$ by

$$\log[p(\mathbf{X}_T|\lambda_c)] = \frac{1}{T}\sum_{t=1}^{T}\log[p(\mathbf{X}_t|\lambda_c)],$$

where the t subscript is used to indicate the speech features associated with the t^{th} frame, T is the total number of frames, and the $1/T$ term is used to normalize to account for varying utterance durations. The likelihood that the speech file is from a speaker other than the claimant (i.e. the background model) is constructed by using a collection of representative speaker models, chosen to "cover" the space of all possible speakers other than the claimant. With a set of B constituent background speaker models $\{\lambda_1, \lambda_2, ..., \lambda_B\}$, the overall background speaker's log-likelihood is computed as

$$\log[p(\mathbf{X}_T|\lambda_{bkgd})] = \log[\frac{1}{B}\sum_{b=1}^{B}p(\mathbf{X}_T|\lambda_b)],$$

where $p(\mathbf{X}_T|\lambda_b)$ is calculated as above. Neglecting the $1/B$ factor, $p(\mathbf{X}_T|\lambda_{bkgd})$ is the joint pdf that the test utterance comes from one of the constituent background speakers, assuming equally likely background speakers.

The motivation for using a ratio of the pdf's of the claimant speaker and the

background speaker, as opposed to just using the claimant speaker's pdf as might seem appropriate initially, is that this normalization has been found to help minimize the nonspeaker-related variations in the test utterance scores[11]. Because the claimant model's scores are likely to be effected in the same manner as the background model's scores, dividing the claimant model pdf by that of the background helps to stabilize the system. This results in a more accurate system that requires less calibration (varying θ) for different environments.

1.3 Mel-Cepstrum and Mel-filter Energy Speech Features

The most common feature vector used in speech systems is that of the mel-cepstrum[8]. For this feature type, every 10 ms the speech signal is windowed by a Hamming window of duration 20 ms to produce a short time speech segment $x_w[n]$. The Discrete Fourier Transform (DFT) of this segment is then calculated and the magnitude is taken, discarding the phase which research has shown to have limited importance in speech. The resulting DFT magnitude $X[m,k]$, where m denotes the frame number and k denotes the frequency sample, is passed through a bank of triangular filters known as a *mel-scale filterbank*, as shown in figure 1-4. The filterbank used is assumed to roughly approximate the critical band filtering that research has suggested occurs in the human auditory system within the outer stage of processing. It is important to note that the Mel-filters do not convolutionally process the DFT magnitude, but rather effectively window the data, weighting each element of $X[m,k]$ by an associated weight provided by the envelop of the Mel-filters. These filters (windows) are linear in their bandwidths from 0 to 1 kHz and are logarithmic above 1 kHz. The centers of the filters follow a uniform 100 Hz Mel-scale spacing and the bandwidths are set such that the lower and upper passband frequencies lie on the center frequencies of the adjacent filters, giving equal bandwidths on the Mel-scale but increasing bandwidths on the linear frequency scale. The number of filters is typically chosen

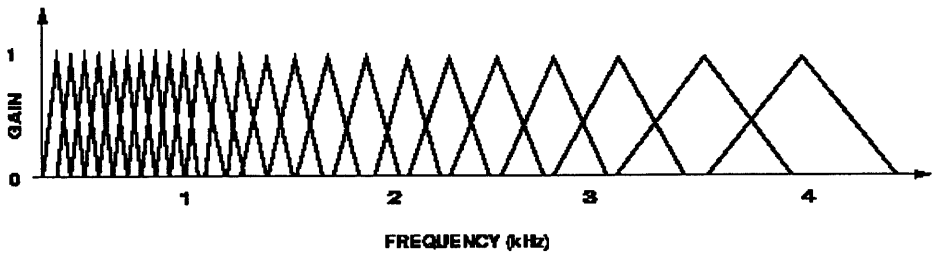

Figure 1-4: Triangular Mel-Scale Filterbank.

to cover the signal bandwidth $[0, f_s]$ Hz, where f_s is the sampling frequency. In most cases the speech being used is telephone speech, in which case $f_s = 8$ kHz and there are 24 filters.

Assuming that there are $\mathcal{N}$ Mel-filters and that $M_l[k]$ represents the values of the l^{th} filter, the linear *mel-filter energy (MFE)* at the output of the l^{th} filter is given by

$$\mathcal{M}_{lin}[m, l] = \frac{1}{A_l} \sum_{k=L_l}^{U_l} M_l[k] X[m, k]$$

where L_l and U_l are the lower and upper frequencies of the l^{th} filter, respectively. In this expression A_l is a normalizing factor used to account for the varying mel-filter bandwidths and is defined as

$$A_l = \sum_{k=L_l}^{U_l} M_l[k].$$

The linear mel-filter energies $\mathcal{M}_{lin}[m, l]$ from the outputs of all $\mathcal{N}$ of the mel-filters gives a reduced representation of the spectral characteristics of the m^{th} speech frame and can be used as a feature vector for speech recognition system. Various techniques discussed later in this thesis operate in this domain.

Due to various advantages involving homomorphic filtering and its ability to linearly filter (lifter) in the frequency domain distortions convolved into the speech in the time-domain as well as certain decorrelating properties of the resulting feature elements, it is common to further process the linear mel-filter energies to produce mel-cepstrum features. The mel-cepstrum features $\mathcal{M}_{cep}[k]$ are computed by taking

the logarithm of the linear mel-filter energy features and taking their inverse Fourier transform:

$$\mathcal{M}_{cep}[m,k] = \frac{1}{\mathcal{N}} \sum_{l=1}^{\mathcal{N}} \log(\mathcal{M}_{lin}[m,l]) e^{j(2\pi/\mathcal{N})lk}.$$

The decorrelating property is due to the high amount of statistical independence resulting from the mathematical similarity of the combination of the log and DFT^{-1} operations to the optimal orthogonalizing Karhunen-Loeve transformation[8]. In addition, the mel-cepstrum representation has been shown in practice to generally give better performance than other feature types in speech related tasks.

1.4 The Speaker Mismatch Problem in Speaker Verification

As discussed in section 1.1, speaker recognition and verification tasks require two stages or processing: training and testing. Both of these require speech and in typical scenarios the speech used for training will be different than the speech used in testing, both in terms of the utterance itself as well as the channel environment in which the speech is produced and recorded. A representative example might be in a banking scenario where a person goes to his or her local bank to open a checking account and in the process has their voice recorded for account access control purposes later. When he or she goes to Hawaii sometime thereafter and tries to withdraw some money from the branch office there, the person will perhaps state some phrase for identification purposes into a recording device quite different from the one used to originally record speech used in making their speech model. Although many channel compensation techniques exist for the purpose of removing the influence of the channel from the features derived from a given utterance, these techniques are not perfect and are often unable to handle all the various complex linear and nonlinear noise processes embedded into the speech signal. Due to such insufficiencies in our available compensation techniques, the speaker models will usually contain characteristics not only derived

from the nature of the speaker's voice, but also from the spectral characteristics of the noise and distortion in the speech, resulting in a *mismatch condition* between the speaker model and a test utterance that reflects more of the environment than the actual underlying speech. Work in speaker identification and recognition tasks has shown that an existence in such a mismatch condition between training and testing speech can often drastically reduce performance. It is very interesting to note that in most cases a mismatch involving clean training data and corrupted testing data, or vice-versa, will perform worse than a scenario where both training and testing data is corrupted. This shows how significant the presence of the channel in the speech features and models can be. This thesis looks into some methods, both existing and new, to help alleviate the performance degrading effect of a mismatch condition between training and testing data.

1.5 Problem Addressed

In this thesis, speaker verification within the setting of the TSID (Tactical Speaker ID) corpus, a collection of realistic recordings made of military personal at Ft. Bragg during spring 1997, is investigated. In total there were 35 speakers involved and recordings were made of soldiers reading sentences, digits, and map directions over a variety of very noisy and low bandwidth wireless radio channels. In this thesis these recordings are referred to as *dirty* speech files. For each dirty recording a low noise and high bandwidth reference recording was made simultaneously at the location of the transmitter with a microphone. These are referred to as *clean* speech files. In figures 1-5 and 1-6 are shown the time and frequency domain representations for the clean and dirty recordings of a representative utterance in the corpus. As may be seen from these plots, the waveform is subject to various types of nonlinear distortion as well as background ambient noise.

As would be expected given the discussion of the training/testing mismatch condition above, some verification tasks using both clean and dirty data for model training and for testing has shown that the clean/clean and dirty/dirty cases do signif-

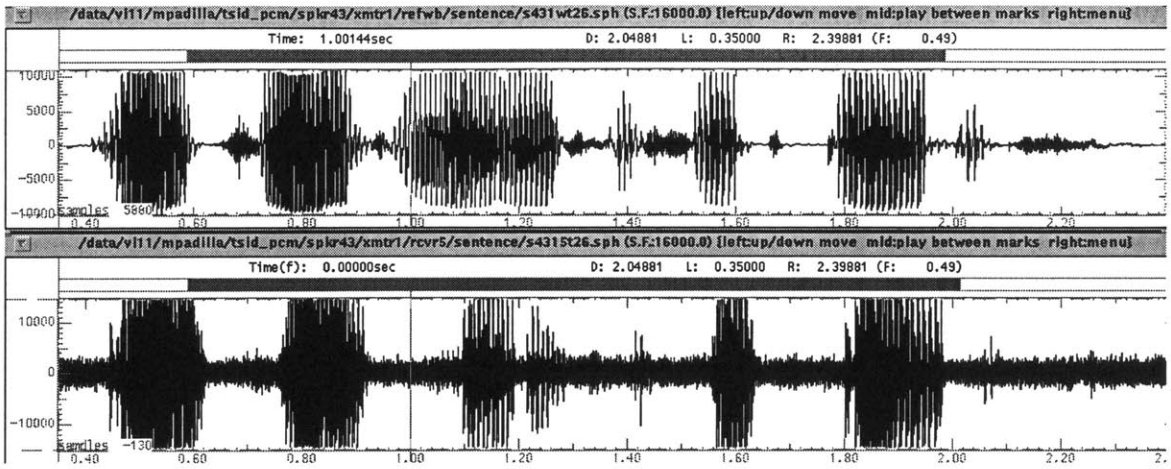

Figure 1-5: Representative "clean" (top) and "dirty" (bottom) speech time-domain waveforms from the TSID corpus. Both are from the same utterance.

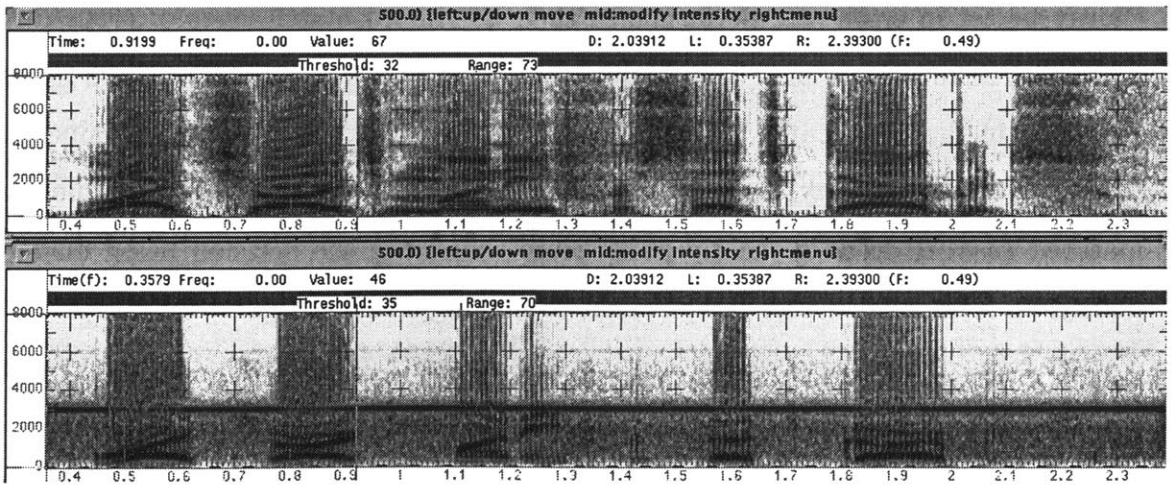

Figure 1-6: Representative "clean" (top) and "dirty" (bottom) speech spectrograms from the TSID corpus. Both are from the same utterance.

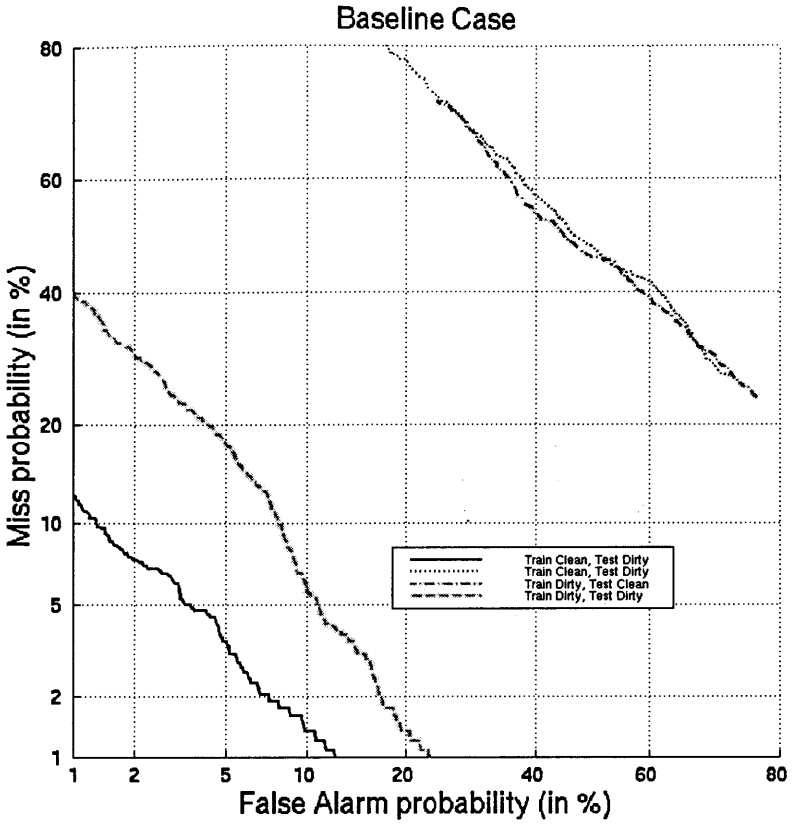

Figure 1-7: DET curve showing error rate performance for baseline tests on TSID corpus.

icantly better in error rate performance than do the mismatch cases of clean/dirty and dirty/clean, which have been found to perform nearly identically. These results are given in the DET curve in figure 1-7. It is seen that while the EER values for the clean/clean and dirty/dirty scenarios are approximately 4% and 9%, for the mismatched cases it drops to nearly 50%. Note that since the decision being made is binary, an EER rate of 50% is equivalent to making a random guess for every test.

The focus of this thesis research is to study the effect of the mismatch condition on this database (training on clean and testing on dirty, or vice-versa) and to investigate ways of improving performance via the application of concepts in *missing feature theory* and *spectral subtraction.* missing feature theory recognizes the fact that due to noise it is possible that some speech features could be corrupted to the degree that they no longer contain accessible speech information and that their inclusion in the statistical scoring mechanism could reduce performance. This requires the detection

of which features are missing and then their removal from scoring, done in such a way as to be consistent with the existing GMM model. This technique, known as missing feature compensation, is investigated along with the ability to accurately predict missing features. In addition, this thesis proposes a new manner to deal with the missing feature problem, *time-frequency linear minimum mean-square error (MMSE) feature estimation*, in which the missing feature is detected, estimated, and then restored. This system is described along with some experimental results. Finally, a technique for subtracting out additive noise effects, known as spectral subtraction, is evaluated as is a new variation on this technique that is proposed in this thesis, to be termed *soft* $|DFT|$ *spectral subtraction.* In this thesis each of these techniques are examined individually and in various combinations. In order to be able to control the evaluation and development of these techniques, throughout most of this thesis the clean TSID data files have been corrupted with known artificial additive noise. It is hoped that an understanding of the techniques explored in this thesis can be most readily achieved by first studying them in a controlled additive noise environment, making the transition later to unknown noise sources, such as the dirty TSID files, more feasible.

1.6 Contribution of Thesis

This thesis contributes to the body of knowledge in speaker recognition, specifically the train/test mismatch problem, in a number of ways. First, the performance of several established channel compensation techniques are applied to the clean/dirty verification problem in the context of the TSID database, thus extending our knowledge of how well these techniques can perform. Second, this thesis investigates the application of spectral subtraction in the $|DFT|$ domain, in contrast to the more common linear mel-filter energy domain, for the purpose of speech feature enhancement. It is demonstrated that the application of spectral subtraction in the $|DFT|$ domain is greatly superior for the type of additive noise channel being considered. Third, the application of missing feature theory, namely missing feature compensa-

tion, is studied and shown to improve performance over the baseline case for certain noise levels. In the process of this study, it is shown that the detection of missing features can be done non-ideally and still produce results highly correlated with an ideal detector. Forth, many of the channel compensation and missing feature methods are combined to study recognition performance when these systems are applied in series. It is found that the combination of $|DFT|$ domain spectral subtraction and missing feature compensation performs very well, outperforming the baseline case and all other methods examined in this thesis. Finally, the concept of missing feature estimation and restoration is explored. A linear minimum mean-squared error missing feature estimator, that operates in both the time and frequency domain, is derived and applied to both pure vowels as well as a TSID clean/dirty verification trial. While it does not improve performance in the verification trial, very promising results regarding the pre- and post-SNR levels for features estimated for the pure vowel indicate promise in the application of this approach.

1.7 Organization of Thesis

This chapter has introduced the general theory and problem of speaker verification, the statistical speaker model and speech features used, and the problem being addressed within this framework along with some of the more important results within this thesis. In Chapter 2 various methods of channel compensation and a method for adding temporal information to the speech feature vectors are discussed along with results using these techniques on the TSID data in the mismatch scenario. Chapter 3 explains in detail the concepts of spectral subtraction and the proposed method of soft $|DFT|$ domain spectral subtraction, followed by a description of missing feature theory and missing feature compensation in chapter 4. Both of these methods are evaluated using speech from the TSID corpus with a mismatch condition imposed through the addition of AWGN to the test data. Chapter 5 investigates what performance benefits are possible when the individual techniques in chapters 3 and 4 are combined in various combinations. The general method of missing feature restora-

tion, which tries to predict missing features rather than discard them as in missing feature compensation, is addressed in Chapter 6 along with the proposed technique of time-frequency linear MMSE missing feature estimation. These techniques are also evaluated with AWGN corrupted TSID speech. Finally, Chapter 7 summarizes the results presented in the thesis and suggests possible areas of future research.

Chapter 2

Channel Compensation and Techniques for Robust Speaker Recognition

The overall goal of speech feature selection in automatic speaker recognition is to extract from a given speech waveform the most linguistically related signal components, while hopefully omitting unwanted elements such as noise and distortion imparted into the signal via both the channel and the speech apparatus itself. These degradations may be additive, linearly convolutive, or nonlinear in nature and may contain modulation frequencies in part of or perhaps across the entire bandwidth of the speech signal. This is especially important in cases where speaker models are trained and tested in different environments, since leaving the channel characteristics in the training speech will bias the speaker models in a way that will be uncompensated for in the testing speech. In essence, the speaker models will reflect not just characteristics of the speaker, but also those of the microphone, recording equipment, etc., resulting in a mismatch condition brought about not by the underlying speech, which is important, but rather by the environment in which the speech was recorded.

The following sections describe two methods for removing channel distortion from the speech features, cepstral mean subtraction and RASTA, as well as a method used to make the features somewhat more robust by including temporal feature informa-

tion. First the methods are described, followed by experimental results using these techniques in clean/dirty verification tasks.

2.1 Cepstral Mean Subtraction

Cepstral mean subtraction (CMS) attempts to remove from cepstral coefficient speech features convolutional distortion from a linear time-invariant channel[8]. Consider a scenario where a clean speech signal $x[n]$ is passed through an LTI system $h[n]$ producing a filtered output $y[n]$. Let this output $y[n]$ be windowed by a time-limited function $w[n]$ in order to extract segments of the signal for short time frequency analysis, producing the sequence $y_w[m,n]$, where m denotes the short time segment and n denotes time

$$y_w[m,n] = (h[n] * x[n])w[m,n].$$

Given $y_w[m,n]$, we would like to be able to recover $x[n]$ without having to first estimate the system transfer function $h[n]$. This is known as *blind deconvolution.* Continuing from the above expression, if we assume that the nonzero duration of $w[m,n]$ is long and relatively constant over the duration of the channel response $h[n]$, then the following approximation may be made in the Fourier domain

$$\begin{aligned} F\{y_w[m,n]\} &= F\{(h[n] * x[n])w[m,n]\} \\ &\approx F\{(x[n]w[m,n]) * h[n]\} \\ &= X_w[m,k]H[k] \end{aligned}$$

which, for a fixed discrete frequency value k, is a time-trajectory in the parameter m. Applying the log operator then gives

$$\begin{aligned} \log\{Y_w[m,k]\} &= \log\{X_w[m,k]H[k]\} \\ &= \log\{X_w[m,k]\} + \log\{H[k]\}. \end{aligned}$$

We see then that each k indexed time-trajectory is the sum of a term that varies with

m given by $\log\{X_w[m,k]\}$ and a constant term given by $\log\{H[k]\}$, representing the channel. Since this channel term is a constant (DC), it may be removed by subtracting the mean value from each time-trajectory. Because typical speech has zero DC, this may be done without disturbing the speech term. This may be done by homomorphic liftering each of time-trajectories in the cepstral domain

$$\begin{aligned}\hat{c}[m,k] &= F^{-1}(\log\{X_w[m,k]\} + \log\{H[k]\}) \\ &= \hat{x_m}[m,k] + \hat{h}[k] \\ &= \hat{x_m}[m,k] + h_k\delta[m].\end{aligned}$$

Applying a cepstral lifter $l[n] = 1$ everywhere except for 0 at the origin will extract the $h_k\delta[m]$ function representing the channel.

2.2 RelAtive SpecTrA

2.2.1 Introduction and Basic System

Research has found that, much like the other sensory systems in the body, the auditory system tends to be sensitive to relative changes in an input signal, rather than absolute levels. Along these lines, there is evidence that the auditory system does not value signal components of all modulation frequencies equally [8]. Rather, there seems to be a peak in sensitivity to modulation frequencies around 4 Hz, with the relative sensitivity gradually decreasing as the frequency is decreased or increased. This modulation frequency of 4 Hz is sometimes referred to as the "syllabic rate". In addition, experiments conducted by van Vuuren and Hermansky[15] suggest that modulation frequencies below 0.1 Hz and above 16.0 Hz do not tend to be very relevant in determining the various speaker scores in automatic speaker recognition. These findings were further supported by Summerfield et al.[13] who showed that the perception of "speech-like" sounds depends strongly on the spectral difference between adjacent sounds, i.e. the temporal fluctuation of spectral components.

RelAtive SpecTrA (RASTA)[4] makes use of these findings by homomorphically filtering (liftering) the time-domain trajectories of the logarithmic cepstral speech features by an IIR bandpass filter with corner frequencies of about 0.26 Hz and 12.8 Hz and a peak at approximately 4 Hz (at the syllabic rate), with sharp notches at 28.9 Hz and 50 Hz. The transfer function for this filter is given by

$$H(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

The high-pass denominator portion of this filter is to alleviate the convolutional noise introduced by the channel, while the low-pass portion is added to help smooth out the fast "instantaneous" spectral changes present in the short-term spectral estimates. The associated frequency and impulse responses are given in figures 2-1 and 2-2.

It is thus seen that RASTA suppresses speech feature components with modulation frequencies outside $H(z)$'s passband while keeping those seen as being most

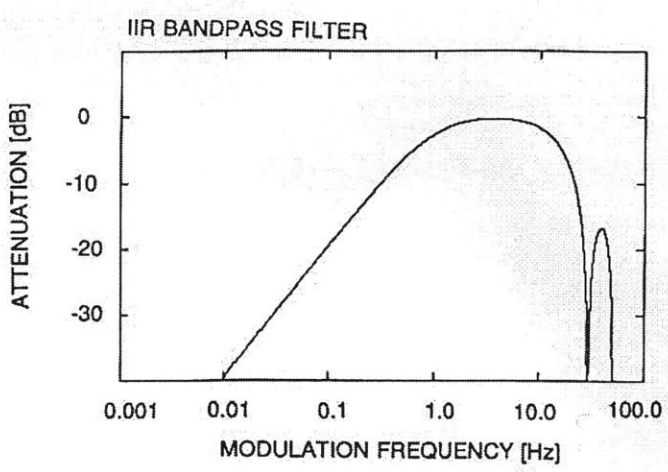

Figure 2-1: The frequency response of the RASTA filter $H(\omega)$.

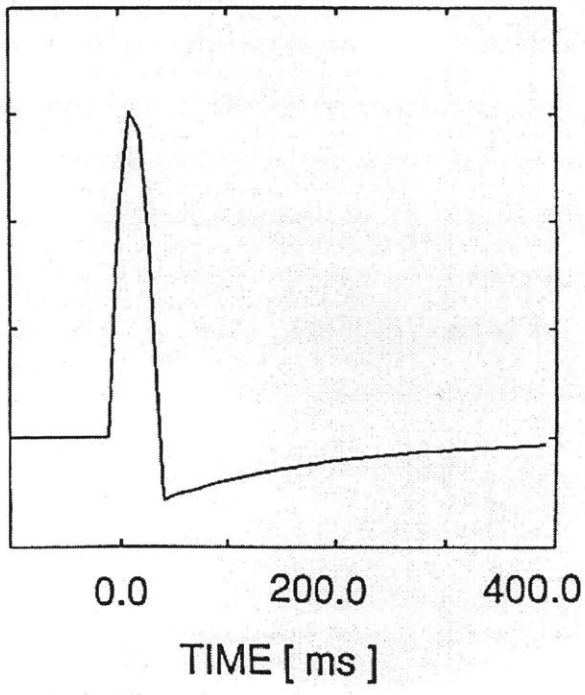

Figure 2-2: The impulse response of the RASTA filter h[n].

relevant to automatic speaker recognition. Due to the spectral nulling at DC and other low frequencies below 0.26 Hz, we see that RASTA filtering will remove any time-invariant, similar to cepstral mean subtraction, as well as slowly time-varying linearly convolutive channel effects for each separate speech feature time-trajectory. The result is a reduction in the sensitivity of the new spectral feature estimate to slow and overly rapid variations in the short-term spectrum, brought about by static or slowly varying channels, noise, and unnecessary speech content.

2.2.2 Lin-Log RASTA

When performed as described above on cepstral features in the logarithmic domain, RASTA processing is most effective in diminishing the effects of time-invariant or slowly time-varying convolutional channel effects that are purely additive or nearly so (since an LTI channel will map to a linear additive term in the log spectral or cepstral domain). However, uncorrelated additive noise components that are additive in the log-magnitude domain, prior to the log operation, become signal dependent after the logarithmic operation is performed. To see this, consider in the frequency domain the output $Y(\omega)$ of a linear system $H(\omega)$ that is excited by a speech signal $S(w)$ and is corrupted by additive noise $N(\omega)$:

$$\begin{aligned} Y(\omega) &= H(\omega)S(\omega) + N(\omega) \\ \log Y(\omega) &= \log\left[H(\omega)S(\omega) + N(\omega)\right] \\ &= \log\left\{S(\omega)[H(\omega) + N(\omega)/S(\omega)]\right\} \\ &= \log S(\omega) + \log\left[H(\omega) + N(\omega)/S(\omega)\right] \end{aligned}$$

It is observed that by the $N(\omega)/S(\omega)$ term the noise is signal dependent and hence nonlinear, making RASTA processing by a fixed linear IIR filter as before not entirely appropriate for its removal from the speech signal.

In order to make RASTA processing more amenable to the removal of this sort of

uncorrelated additive noise, a family of alternative nonlinear transforms are employed instead of the standard logarithm. These nonlinear transforms have the property that they tend to be linear-like for small spectral values while log-like for large ones. The transform is parameterized by the single positive constant J and is given by

$$Y = \log(1 + JX)$$

where X is the linear-domain speech feature and Y is the nonlinear transform of this feature. To gain some insight into the nature of this operation, it helps to rewrite the above expression slightly:

$$\begin{aligned} Y &= \log(1 + JX) \\ &= \log[J(1/J + X)] \\ &= \log(J) + \log(1/J + X). \end{aligned}$$

Upon the application of the RASTA filter $h[n]$, the first term (a constant) is removed due to the high-pass portion of the filter. It is then seen that lin-log RASTA may be seen as a form of noise masking in which a fixed amount of additive noise is added prior to the RASTA processing in the logarithmic domain. This is somewhat similar to a technique employed in generalized spectral subtraction, which is described in chapter 3, in which additive noise is added to help mask musical tones[8].

To return to the linear domain, the following inverse is applied

$$X = \frac{e^Y - 1}{J}.$$

2.3 Delta Cepstral Coefficients

Delta cepstral coefficients (DCC) are used to impart temporal information into the speech feature vector. In computing the coefficients, a low order polynomial is fit to to the features within a given window length over successive frames for each feature trajectory. The parameters of the derivative of the polynomial are used as an extra feature, appended to the end of the usual feature types discussed in section 1.3. In this sense, the use of this technique allows the speech features to contain information regarding the rate of change of formants and other spectral characteristics, possibly unique to a given speaker [12].

2.4 Results with TSID Corpus

To evaluate how well these techniques perform on real data, each of them was applied to the train clean/test dirty mismatch scenario with the TSID corpus. In this case, the actual narrowband noisy cellular data is used for the dirty test data. In all of the tests done in this section, the speech features used were cepstral coefficients and the sampling rate was 8 khz, resulting in 24 features. GMM speaker models were of order 1024.

As a means of comparison to what is possible in an ideal matched clean/clean case without any equalization applied (trials performed without any form of channel compensation of other noise management technique will hereafter be referred to as the *baseline* case), it was found that the corresponding EER was 4.2%. In contrast, it was found that the baseline mismatch case of train clean/test dirty [1] had an EER of approximately 50%, signifying that in the baseline mismatch case for the TSID data the system is essentially flipping a coin to make its accept/reject decisions.

In applying the techniques of CMS, RASTA, and DCC, both individually and in combination, the results in table 2.4 were found.

EQUALIZATION	EER
Baseline + CMS + DCC	23%
Baseline + CMS + RASTA	28%
Baseline + CMS	28%
Baseline + RASTA	34%
Baseline + DCC	43%
Baseline	49%

Table 2.1: Results applying various channel compensation techniques to TSID data for train clean/test dirty case.

The corresponding DET curve is given in figure 2-3.

[1] Results throughout this research have shown that the performance in many cases of the train clean/test dirty case is nearly identical to that of the train dirty/test clean case. Due to this observation and the fact that the most likely scenario is clean data during training and dirty data during testing, this thesis focuses on the train clean/test dirty case.

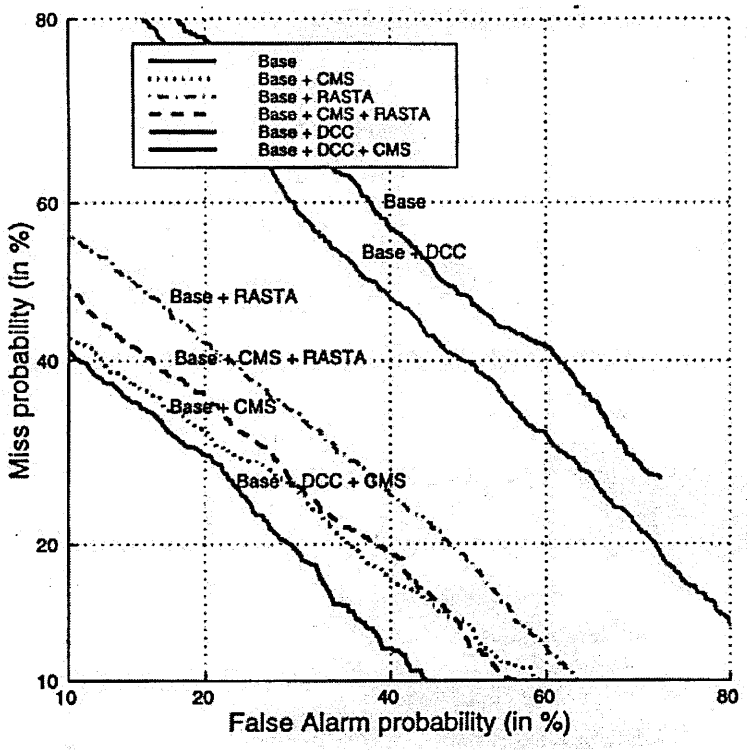

Figure 2-3: Comparison of various equalization techniques.

Important points to note are that the Baseline + RASTA + CMS case produced an EER of 28%, which is the same as the Baseline + CMS case, without RASTA being applied. As discussed in the introductory part of this chapter, RASTA has been developed for the purpose of removing slowly varying convolutional distortion, hence its removal of small quefrencies[2] from 0 to 0.1 Hz. CMS, on the other hand, is only able to remove strictly time-invariant convolutional distortion through the liftering out of the 0^{th} cepstral coefficient for each temporal trajectory. The fact that the EER values for the systems of CMS + RASTA and CMS alone perform the same at 28% indicates that for the TSID corpus that the most corruptive convolutional distortion is likely to be of a time-invariant nature, hence the lack of additional improvement from RASTA when CMS is already applied. It is also observed that DCC alone does not help performance very much, while adding DCC on top of CMS improves performance slightly from 28% to 23%, producing the best performing system that

[2]Frequency in the cepstral domain is frequently referred to as "quefrency", noting the reversed role played by time and frequency in this domain. This also explains the the motivation for the terminology "cepstral" (spectral) and "liftering" (filtering).

improves the baseline EER by 26%.

Chapter 3

Spectral Subtraction

This chapter investigates a general methodology to deal with the corruption of speech features with additive noise. It is a well established and intuitively reasonable fact that as speech is degraded by ambient background noise that the resulting automatic speaker recognition performance decreases dramatically. One approach to alleviating this effect is to devise speech features that are sufficiently robust to the noise types and levels so as to make estimation of the noise characteristics unnecessary. The approach discussed in this chapter, *spectral subtraction (SS)*, makes the assumption that the type and level of noise affecting the speech is amenable to being effectively estimated attempts to derive a spectral characterization of the additive noise process effecting the speech such that it may be subtracted from the spectral magnitude of the corrupted speech waveform. While this operation is typically done in the $|DFT|$ domain, applications in the speech recognition and verification for the purpose of cleaning up speech features have focused on performing the spectral subtraction in the linear Mel-filter energy domain. This thesis proposes that it be done one stage earlier in the calculation of the speech features at the $|DFT|$ stage prior to the application of the Mel-filters. This provides a "softer" form of spectral subtraction that will be shown to be superior to traditional spectral subtraction. This chapter concludes by reporting several results obtained when these techniques were applied to the TSID corpus in a clean/dirty mismatch scenario, the dirty data being a result of additive white Gaussian noise.

3.1 Mel-Filter Energy Domain Spectral Subtraction

First we will consider the spectral subtraction in the domain in which is has typically been applied when the goal is to enhance speech features, in the linear mel-filter energy domain.

Consider a speech signal $x[n]$ that has been corrupted by additive stationary ambient noise $a[n]$, producing a noisy speech signal $y[n]$

$$y[n] = x[n] + a[n],$$

the spectral magnitude of which is

$$\begin{aligned} |Y(\omega)|^2 &= |X(\omega) + A(\omega)|^2 \\ &= [X(\omega) + A(\omega)]^*[X(w) + A(\omega)] \\ &= |X(\omega)|^2 + |A(\omega)|^2 + X^*(\omega)A(\omega) + X(\omega)A^*(\omega). \end{aligned}$$

If we make the reasonable assumption that $x[n]$ and $a[n]$ are uncorrelated wide-sense stationary (WSS) processes, then taking expectations gives us

$$\begin{aligned} E[|Y(\omega)|^2] &= E[|X(\omega)|^2 + |A(\omega)|^2 + X^*(\omega)A(\omega) + X(\omega)A^*(\omega)] \\ &= E[|X(\omega)|^2] + E[|A(\omega)|^2] + E[X^*(\omega)A(\omega)] + E[X(\omega)A^*(\omega)] \\ &= E[|X(\omega)|^2] + E[|A(\omega)|^2] + E[X^*(\omega)]E[A(\omega)] + E[X(\omega)]E[A^*(\omega)] \\ &= E[|X(\omega)|^2] + E[|A(\omega)|^2] \end{aligned}$$

since the complex spectra of both the speech and the ambient noise will typically be zero mean. If the spectral magnitude $|Y(\omega)|^2$ is then input to the mel-filterbank, composed of filters $M_l[k]$, then *on the average* the resulting linear mel-filter energy

features $\mathcal{M}_{lin}[m,l]$ will be composed of a signal component and an additive noise component

$$\mathcal{M}_{lin}[m,l] = \mathcal{M}_{true}[m,l] + \mathcal{N}_{lin}[m,l],$$

where $\mathcal{M}_{true}[m,l]$ denotes the true underlying linear mel-filter energy that we would like to recover. If the non-speech parts of the utterance which only contain the ambient noise are isolated, then an estimate of the power spectral density $E[|A(w)|^2]$ can be made and similarly processed by the mel-filterbank to produce an estimate of the linear mel-filter energies for the noise, $\hat{\mathcal{N}}_{lin}[m,l]$. The true value of the signal component of the corrupted linear mel-filter energy features is then estimated as

$$\hat{\mathcal{M}}_{true}[m,l] = \mathcal{M}_{lin}[m,l] - \hat{\mathcal{N}}_{lin}[m,l].$$

It is thus seen that while instantaneously we may not be able to subtract out the linear mel-filter energy of the noise and recover the linear mel-filter energy of the clean speech signal, given that a good estimate of the spectral magnitude of the ambient noise is available and that the statistical conditions are met, we may approximately do so in an average sense[1].

Given that the estimate for the linear mel-filter energy of the noise will often *instantaneously* exceed that of the noisy speech signal, practical frame-by-frame implementations of spectral subtraction are usually given as

$$\hat{\mathcal{M}}_{true}[m,l] = \begin{cases} \mathcal{M}_{lin}[m,l] - \hat{\mathcal{N}}_{lin}[m,l], & \text{if } \mathcal{M}_{lin}[m,l] > \hat{\mathcal{N}}_{lin}[m,l] \\ 0, & \text{otherwise.} \end{cases}$$

This is done to avoid physically impossible negative energy estimates.

Due to issues involving artificially created musical tones in re-synthesized speech when spectral subtraction is used in the $|DFT|$ domain, to be studied next, the

[1]This distinction between instantaneous and average results is an important one which will have implications later

operation of spectral subtraction in the linear mel-filter energy domain is typically parameterized by two parameters, α and β as spectral subtraction applied in the $|DFT|$ domain does. Although this motivation is not directly relevant to the linear mel-filter energies, a similar formulation has been adapted in this domain as well. The relationship for this perceptually motivated form of spectral subtraction, referred to as *generalized spectral subtraction*[1], is given by

$$\mathcal{D}[m,l] = \mathcal{M}_{lin}[m,l] - \alpha\hat{\mathcal{N}}_{lin}[m,l]$$

and

$$\hat{\mathcal{M}}_{true}[m,l] = \begin{cases} \mathcal{D}[m,l], & \text{if } \mathcal{D}[m,l] > \beta\hat{\mathcal{N}}_{lin}[m,l] \\ \beta\hat{\mathcal{N}}_{lin}[m,l], & \text{otherwise,} \end{cases}$$

where $\alpha \geq 1$ and $0 < \beta \leq 1$. α is used to overestimate the linear mel-filter energy of the noise, and β determines the level of the spectral flooring as a fraction of the estimated noise linear mel-filter energy. The motivation for the use of these parameter will become clearer in the following section.

3.2 Soft $|DFT|$ Domain Spectral Subtraction

In the previous section the development of generalized spectral subtraction, as applied to the linear mel-filter energies, was discussed. While this technique does have the attractive characteristic of removing the additive noise in an average sense, as discussed above it instantaneously may result in a physically impossible negative clean mel-filter energy estimate. To avoid this scenario a nonlinearity is applied that essentially zeros any clean energy estimates falling below the threshold $\beta\hat{\mathcal{N}}_{lin}[m,l]$.

While this may seem to be the most appropriate thing to do in this situation, it raises the issue of there being possible additional performance reductions resulting from this highly nonlinear flooring operation being applied to instantaneous values,

rather than to averages. If the values involved were long-term averages of the signal and noise spectral densities, then perhaps this method of handling negative estimates would be always valid since at least the resulting estimate after the flooring would always be closer to the true value. Instantaneously however, the interpretation needs to be a bit more careful. For example, even if the noise is additive white Gaussian noise and hence has a flat power spectral density, at a given frequency its instantaneous power may actually be very low. In this case the spectral subtraction algorithm is overcompensating for the noise and, if the instantaneous signal power is low enough, may result in the flooring to take effect and the estimate being set to a near zero value, $\beta\hat{\mathcal{N}}_{lin}[m,l]$. When this occurs a potentially clean mel-filter energy has been replaced with the near zero value since the instantaneous noise power density was less than its average, a potentially very common occurrence. It is thus seen that the application of the flooring operation, while necessary to avoid physically impossible negative power estimates, will likely instantaneously result in many poor resulting estimates for the clean mel-filter energy features.

It is this consideration that makes the application of spectral subtraction-based feature enhancement in the $|DFT|$ domain a potentially more attractive option than the way it is currently being pursued, i.e., in the linear mel-filter energy domain[1]. Because the nonlinear flooring operation is now applied to the $|DFT|$ values, prior to the mel-filterbank, the error component of each spectrally subtracted value $X[m,k]$ may be decreased through the averaging that occurs as these values are put through a weighted sum via the mel-filters $M_l[k]$. To make this clearer, consider the following two equations that show the resulting linear mel-filter energy $\mathcal{M}_{lin}[m,l]$ when spectral subtraction, symbolized by the operator $\mathcal{L}$, are applied in the $|DFT|$ domain and in the linear mel-filter energy domain:

$$\mathcal{M}_{DFT}[m,l] = \sum_k M_l[k]\mathcal{L}\{X[m,k] - \hat{N}_{DFT}[m,k]\}$$

and

$$\mathcal{M}_{MEL}[m,l] = \mathcal{L}[\{\sum_{k} M_l[m,k]X[m,k]\} - \hat{N}_{MEL}[m,k]],$$

where $\mathcal{M}_{DFT}[m,l]$ is the linear mel-filter energy when spectral subtraction is applied in the $|DFT|$ domain, $\mathcal{M}_{MEL}[m,l]$ is the linear mel-filter energy when spectral subtraction is applied in the linear mel-filter domain, $\hat{N}_{DFT}[m,k]$ is the noise spectral estimate in the $|DFT|$ domain, and $\hat{N}_{MEL}[m,k]$ is the noise spectral estimate in the linear mel-filter energy domain. It is seen in this these two espressions that while $\mathcal{M}_{MEL}[m,l]$ is exposed to a single nonlinearity operation $\mathcal{L}$, $\mathcal{M}_{DFT}[m,l]$ is computed by summing the outputs of large number of weighted nonlinear operations $\mathcal{L}$. It is reasonable to think that this may allow for some of the error introduced by the nonlinearity to be averaged out. A study of the fundamental differences between SS in these two domains is needed to better understand any peformance differences.

3.3 Results for Spectral Subtraction

Experiments were run to evaluate the performance of both linear mel-filter energy and $|DFT|$ domain spectral subtraction. These tests were done for the clean/dirty case where, as has been the noise scenario used during this paper, the corruption was AWGN. Because in all of the tests in this thesis it was found that relative performance at all miss and false acceptance probability levels was reflected in the relative EER differences, for clarity, results will be reported in terms of EER[2]. In all of the tests done in this section, the speech features used is scoring are logarithmic mel-filter energies and the sampling rate was 8 khz, resulting in 24 features. 1024 mixtures were used in each GMM speaker model.

The performance for linear mel-filter energy domain SS, shown in figure 3-1, is seen to depend on the level of noise corrupting the speech. For SNR values below approximately 13 dB application of SS in this domain was found to improve performance over the baseline by about 4%, while for higher SNRs performance was hurt, by roughly 5%.

In contrast to this result, we see in figure 3-2 that for spectral subtraction applied in the $|DFT|$ domain that performance is improved substantially *at all SNR levels* over baseline. At high SNRs above about 15 dB the EER is improved by about 5% and this improvement increases by close to 10% at low SNR levels. To emphasize the imrovement this form of SS gives, results for RASTA processing are also included in the figure. It is seen that $|DFT|$ domain SS provides substantially lower SNR than RASTA in an AWGN environment.

[2]Although it has been observed that relative performance differences between two systems as reflected in EER often reflects relative performance at all the different possible threshold levels described by the DET curve, this is not always the case and care must be taken to not make overgeneral conclusions based on EER.

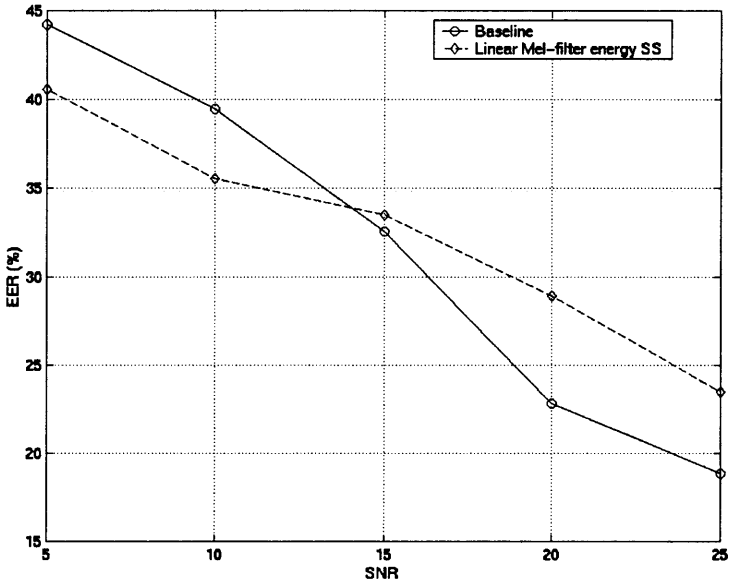

Figure 3-1: Plot showing EER vs. SNR (AWGN) for both the baseline case as well as with linear mel-filter energy domain spectral subtraction. It is seen that the performance relative to the baseline depends on the SNR of the additive noise, gains over baseline occuring at lower SNR levels.

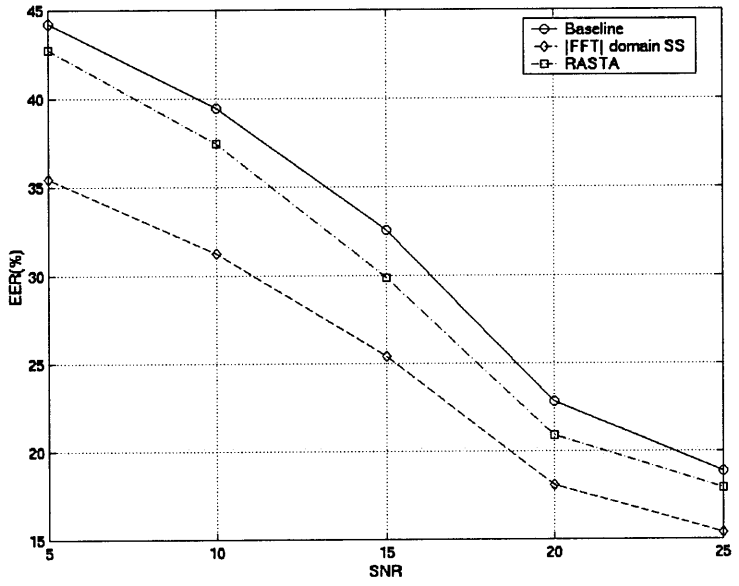

Figure 3-2: Plot showing EER vs. SNR (AWGN) for both the baseline case as well as with $|DFT|$ domain spectral subtraction. The application of spectral subtraction in this domain is seen to improve performance over the baseline case significantly at all SNR levels, going from about a 4% improvement at high SNRs to a near 10% improvement at low SNRs. For comparison, results for RASTA processing are also shown.

Chapter 4

Missing Feature Theory

This chapter investigates *missing feature theory*, which takes a approach different from that seen in chapter 3 in addressing the problem of handling noisy speech. This method attempts to estimate which features are most heavily corrupted and labels them as "lost", to be either removed from scoring or replaced with a new estimates of their true values. This is opposed to being "present", in which case it is estimated that the speaker related information has not been masked out by the noise so as to be unusable. There are two basic techniques related to this concept that will be studied in this thesis. The first of these, known as *missing feature compensation (MFC)* involves dynamically modifying the GMM in order to remove terms associated with the missing features and is the topic of this chapter. The second method is known as *missing feature restoration (MFR)* and will be developed in chapter 6.

This chapter finishes by reporting several results obtained when these techniques are applied to the TSID corpus in a clean/dirty mismatch scenario, the dirty data being a result of additive white Gaussian noise.

4.1 Missing Feature Theory

In any speaker recognition system in practice, the speech being analyzed almost unavoidably is subject to noises that naturally degrade performance. While there are a wide range distortion types that can be encountered, a common model of distortion is additive noise, specifically additive white Gaussian noise (AWGN). While there are many techniques such as RASTA and cepstral mean subtraction have been developed to preprocess the speech and reduce the influence of convolutive noise, these techniques often fail to adequately remove the noise in time-frequency regions dominated by additive noise. While often it is possible to perhaps remove the noise, this may be necessarily accompanied by an amount of speech distortion that nulls out any possible performance gain.

Missing feature theory addresses the topic of corrupted speech features and essentially recognizes the fact that the above mentioned noise reduction/speech distortion trade-off exists and that at times a feature may be too corrupted to be able to use effectively. The idea is that while there may be useful speech information contained within a given speech feature, it is possible that the amount of noise in that time-frequency region is high as to effectively bury any useful speaker dependent spectral information. In this case the inclusion of this highly corrupted feature in the scoring mechanism of the GMM may only worsen performance. The question then becomes how the automatic speech recognition system should handle the missing feature once it is found.

4.2 Missing Feature Compensation

One manner discussed by Lippman in [6] as well as Drygajlo and El-Maliki in [1] to deal with a missing feature is to dynamically adapt the GMM stochastic speaker models to remove it from inclusion in scoring. This method is referred to as missing feature compensation and uses generalized spectral subtraction to classify features as "missing" or "present", rather than as a speech enhancement preprocessor. Recalling

the expression for generalized spectral subtraction in 3.1, a similar relation is now used to detect missing features:

$$D_m(\omega) = |Y_m(\omega)|^2 - \alpha|\hat{A_m}(\omega)|^2$$

and

$$Y_m(\omega) = \begin{cases} \text{“present”}, & \text{if } D_m(\omega) > \beta|\hat{A_m}(\omega)|^2 \\ \text{“missing”}, & \text{otherwise,} \end{cases}$$

where $\alpha \geq 0$ and $0 < \beta \leq 1$.

As discussed in the introduction, the speaker verification system used in this thesis was based on GMM speaker models, which are probability density functions comprised of a linear combination of multi-dimensional Gaussian densities, each of which describes a particular speech “state” of the speaker

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i \Phi_i(\mathbf{X}, \mu_i, \Sigma_i)$$

where p_i is the is the probability of the speaker being in state i, Φ is a multi-dimensional Gaussian pdf for that state with mean μ_i and covariance matrix Σ_i, and M is the number of speech states. Although the off-diaganal terms in Σ_i are typically nonzero, for the order of most GMMs keeping all those terms would lead to an excessive computational burden. To alleviate this problem, Σ_i is often assumed to be diagnol and its off-diagonal terms are forced to 0. Because of this assumption each of the multi-variate Gaussian pdf's Φ may be rewritten as a product of single-variate Gaussian pdf's as follows

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i \prod_{k=1}^{D} \Phi_i(x_k, \mu_{ki}, \sigma_{ki}^2)$$

where D is the dimension of the feature vectors and μ_{ki} and σ_{ji}^2 are the mean and variance of feature element x_k. Given our Spectral Subtraction based missing feature detector, we may label each of our feature elements x_k as being either “present” or

"missing" and hence divide our overall speech feature vector $\mathbf{X} = \{x_1, x_2, ..., x_D\}$ into two sub-vectors, $\mathbf{X}_{present}$ and $\mathbf{X}_{missing}$. With this regrouping our above equation becomes

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D_{present}} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2) \prod_{k=1}^{D_{missing}} \Phi_i(x_k, \mu_{ki}, \sigma_{ki}^2).$$

Missing Feature Compensation then proceeds by removing the second product term representing the sub-vector $\mathbf{X}_{missing}$ and leaving only the data associated with $\mathbf{X}_{present}$

$$p_{mfc}(\mathbf{X}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D_{present}} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2).$$

This new pdf $p_{mfc}(\mathbf{X}|\lambda)$ is then used in place of the full multivariate GMM $p(\mathbf{X}|\lambda)$.

One potential problem with the missing feature compensated speaker model is the fact that this new probability density function is by no means guaranteed to be properly normalized to represent a valid probability space, and in most cases it will not. For example, while it may be true that the original pdf $p(\mathbf{X}|\lambda)$ would integrate over its constituent feature space to 1, there is no reason to think that the same will be true for $p_{mfc}(\mathbf{X}|\lambda)$. While this may appear to be a potentially detrimental flaw with missing feature compensation, it has been found experimentally that in the case of speaker verification that this does not tend to be the case. The reason for this seems to be that the use of a background speaker model to normalize the likelihood scores in speaker verification not only helps normalize for any variations in the channel, but also provides a degree of probability space normalizing in the case of MFC. Several speaker verification tasks were run and the speaker scores versus time were looked at along with the corresponding scores for the background model. As discussed in chapter 1, in speaker verification tests the speaker scores are normalized in the log domain by the corresponding scores of a background model, in other words

$$\Lambda(\mathbf{X}_T) = \log[p(\mathbf{X}_T|\lambda_c)] - \log[p(\mathbf{X}_T|\lambda_{bkgd})].$$

It is hoped that the inclusion of the background model in the expression above will

allow for variations in the speaker score caused by missing feature compensation to be compensated for by similar variations in the background score, such that the difference seen above will remain relatively constant. To investigate this, a controlled experiment was run in which a predetermined feature was declared missing for every frame and the corresponding normalized (by the background model) scores for a given speaker were compared to the same normalized scores for the same test but with no features declared to be missing. This was done with clean data. Plotting these two sets of normalized scores for the case where only the 10^{th} feature was declared missing every frame resulted in the plot seen in figure 4-1. This linear relationship between these two sets of scores demonstrates the fact that the speaker and background scores tend to vary together, and as a result the component of the speaker scores that is attributable to the inclusion or removal of mixtures associated with the present and missing features is normalized by a similar component in the background scores. This is a result of the fact that decisions made regarding present and missing features hold for all speakers, including the background model as well, resulting in similar missing feature compensation in both. It should be noted, however, that this will not hold for any number of features declared missing. As was may be seen in figure 4-3, EER performance will be generally maintained given that the number of missing features declared and removed is below a certain threshold (roughly equal to 15 in figure 4-3), beyond which the loss of speaker dependent spectral information begins to dominate performance.

To investigate in a controlled manner the potential performance benefit that missing feature compensation can offer, a few experiments were done in which additive noise and feature nulling were used in combination with MFC. ditive noise and feature nulling were used in combination with MFC. The first point was the question of how much performance degradation might occur if a single feature was truly missing, in the sense that it was perhaps subject to tonal distortion in its associated spectral range during a given frame. To simulate this condition, a test was run in which the 10^{th} linear mel-energy feature for every frame (out of 24 features) for an otherwise clean/clean case. Every other feature was uncorrupted. The choice of the 10^{th} feature

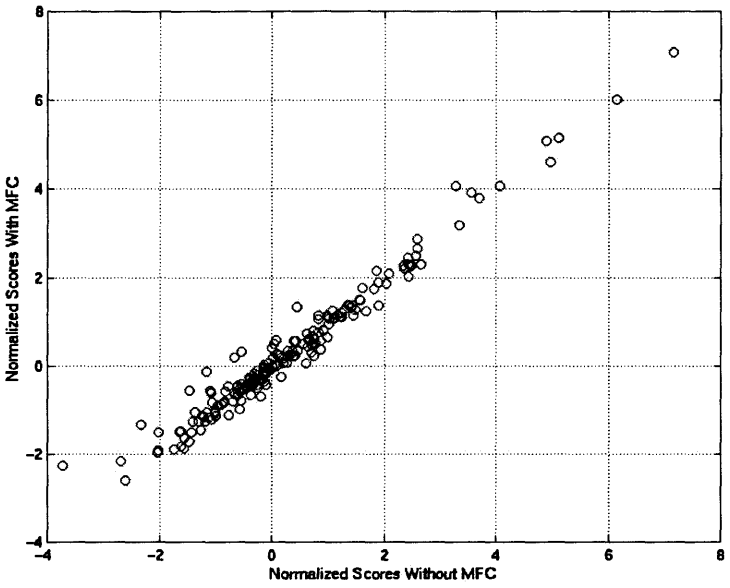

Figure 4-1: This plot shows the background model normalized log probability scores without MFC versus the corresponding values with MFC. The hope is that normalization of speakers' log probability scores with the background model's score will help to properly normalize the probability space in cases where MFC is applied. In this plot, the scores where MFC was applied resulted from removing the 10^{th} feature from every frame. It is seen that corresponding scores with and without MFC tend to be highly correlated, supporting the claim the background model normalization helps correct for improperly normalized pdf's when MFC is applied. Similar results were seen when slightly more features were removed in a controlled fashion as well.

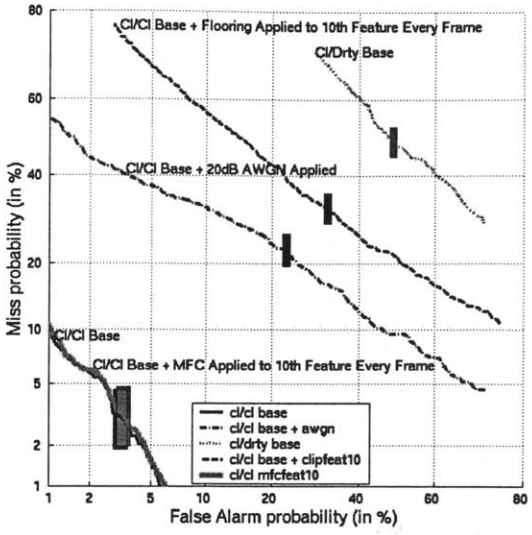

Figure 4-2: Illustration of DET performance with 10^{th} linear mel-filter energy feature floored and with the same feature removed via MFC. MFC is seen to recover the baseline performance. DET curves with 20 dB of additive noise are given for reference.

was made since it has been experimentally shown to represent information in a relatively important spectral band[15]. The resulting EER point was found to be 32%, in comparison to an EER of 3.26% in the uncorrupted clean/clean case. This indicated that even a single feature, if heavily corrupted, could severely impact recognition performance. Given that even a single heavily noise masked feature per frame would have this effect, it is of interest to see the impact that removing that feature from every frame via MFC would have. To this end another test was run in which MFC was applied to the 10^{th} feature in every frame, resulting in an EER of 3.39%, only slightly worse that the ideal clean/clean baseline result. These results suggest that significant performance gain can be potentially realized through accurately detecting, and removing, missing features. The DET curve for these tests is in figure 4-2[1].

One important concept regarding the use of MFC is that with any feature that is removed from scoring, not only is the masking noise being discarded but so is any speech information contained in the feature. It is conceivable that in some cases the removal of a given feature's noise from the scoring mechanism via MFC may

[1]The blocks seen in this figure represent confidence/significance intervals assuming an underlying binomial distribution.

compensate for the associated loss of speech information, but at a certain quantity of removed features the absence of speech information will likely become too great to make further application of MFC beneficial. To investigate this question and develop an understanding of the trade-off involved, another set of experiments were done which involved the corruption of randomly chosen features every frame for the clean test data. The number of corrupted mel-filter energy features was fixed, although the particular features these would be was determined randomly every frame, each frame independent of the others. The corruption was additive white Gaussian noise at 20 dB. One test simply looked at verification performance using this randomly corrupted data. For comparison, another test was done in which for each frame a certain predetermined number of features were chosen randomly to be removed by MFC. In addition, a third trial was run where 20 dB of noise was added to all features and a fixed number of features in each frame were randomly chosen to be removed by MFC. The resulting EER points may be seen in figure 4-3. One very interesting result seen in these results is that the application of MFC has the property of maintaining EER performance up to the removal of about 15 mel-filter energy features, outperforming the case of no MFC up to the removal of about 20 features. It is thus seen that, as expected, the application of MFC in these controlled experiments does help to a certain point, beyond which the noise removal cannot compensated for the loss of necessary speech information to do effective speaker verification and the performance starts to worsen rapidly.

4.3 Missing Feature Detection

One underlying assumption in the work discussed thus far has been that the identity of which features are missing and which are present is known. Not only does this entail an ability to reasonably detect or estimate the amount of corruption in a given feature, but it also requires an assumption about what it means to be "missing". There are many ways of measuring error, some typical definitions being mean-square error and absolute error. In addition, for each possible manner by which the error

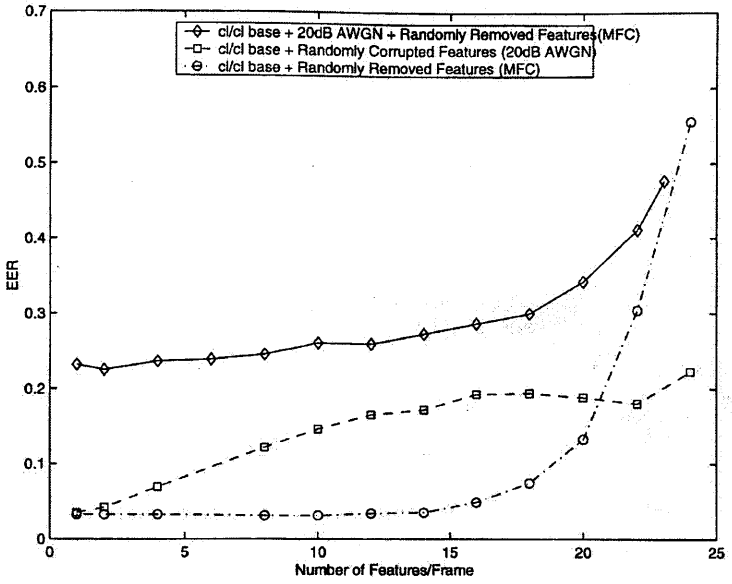

Figure 4-3: Trade-off between corrupted features and removing speech information in MFC. This curve shows EER performance with and without MFC for an increasing number of randomly corrupted or removed mel-energy features.

may be measured, there are a range of possible threshold values that may be chosen to divide between features classified as present and those classified as missing. In this work the assumption made is that error distance is the criterion of importance. As is suggested by Drygajlo and El-Maliki [1], the missing feature detector used in this thesis is generalized spectral subtraction, which in the linear mel-energy domain is given as follows

$$\mathcal{M}_{lin}[m,l] - \alpha\hat{\mathcal{N}}_{lin}[m,l] \underset{\mathcal{M}_{lin}[m,l] \text{ "missing"}}{\overset{\mathcal{M}_{lin}[m,l] \text{ "present"}}{\gtrless}} \beta\hat{\mathcal{N}}_{lin}[m,l],$$

where $\mathcal{M}_{lin}[m,l]$ and $\hat{\mathcal{N}}_{lin}[m,l]$ are the l^{th} corrupted feature and estimate of the noise component of that feature for the m^{th} frame, respectively[2]. Because the motivation for the inclusion of β is that in speech enhancement applications of GSS in the $|DFT|$ domain it partially determines the amount of noise added to the enhanced speech,

[2]Although not strictly true do to nonlinearities in the steps leading to the linear mel-filter energies, the assumption will tacitly be made here that the signal and noise components of the features add to produce the resulting corrupted feature.

in the aim of reducing the amount of perceived musical noise. In the application discussed in this section this motivation for having a nonzero β is gone, and hence β will typically be set to zero since it only contributes noise to the enhanced features.

Given the assumption that the criterion for detecting missing features above is correct, a "*perfect*"[3] missing feature detector is available for cases where artificial additive white noise is added since both the true clean linear mel-energy feature as well as its corrupted version are available. Hence $\hat{N}_{lin}[m,l]$ may be found by finding the difference between the two and then used in the detection formula above. In contrast to this is what will be referred to as the "*nonperfect*" detector, in which case the expressions above still hold but where the noise estimate $\hat{N}_{lin}[m,l]$ is derived from averaging the features that result during silence frames[4]. In all of the experiments discussed in this thesis, this noise estimate $\hat{N}_{lin}[m,l]$ was a result of averaging a very large number of silence frames such that it is very close to the true value. As in the perfect detector, β is normally set to zero.

[3]This detector is perfect given the assumption stated earlier in this section.

[4]Speech/silence decisions necessary for all the various experiments done during this thesis were produced by a program called "Xtalk", written by Doug Reynolds at the MIT Lincoln Laboratory. Xtalk operates by tracking the minimum and maximum energy levels in a given utterance and estimating that frames with an energy level above some percentage of the difference between these values is speech.

4.4 Results for Missing Feature Compensation

One of the first issues addressed with regard to techniques that require missing feature detection is the question of how to calibrate the detector, or rather, what value of α results in the best performance. To answer this question, several initial experiments were run in which various levels of noise from 5 dB to 20 dB were added and enhancement using linear Mel-filter energy domain spectral subtraction was done with α values ranging from 0.5 to 5.0. As before, logarithmic mel-filter energies were the selected feature type for scoring and the sampling rate was 8 khz, resulting in 24 features. GMMs were of order 1024. Performance was judged in terms of EER, and at each noise level the value of $\alpha = 3.0$ was found to result in the best performance. A similar set of tests were done in which the perfect detector was used in conjunction with MFC, in which case $\alpha = 3.0$ was again found to be the optimal value.This is interesting in that a value of $\alpha = 3.0$ suggests that it is preferred to somewhat *overestimate* the amount of energy in the noise during the subtraction. This tends to agree with the result associated with figures 4-2 and 4-3, which suggest that there is indeed a trade off between removing noise and removing speech information and that it is detrimental to try to remove all noise. From these results it may be said that it is only the features that are effected most heavily that should be removed as doing otherwise would potentially remove too much speech information. Based on these results all missing feature detection was performed with values $\alpha = 3.0$ and $\beta = 0.0$.

In any real application of spectral subtraction and MFC, naturally perfect instantaneous knowledge of the noise component of $\mathcal{M}_{lin}[m, l]$, $\mathcal{N}_{lin}[m, l]$ will be unattainable and the best that can be hoped for is an estimate that is close to the average of this process:

$$\hat{\mathcal{N}}_{lin}[m, l] \rightarrow \overline{\mathcal{N}}_{lin}[m, l].$$

In order estimate this average, features resulting from nonspeech frames may be averaged, either globally if the noise is stationary or adaptively if the noise is nonstationary. One question of interest then is how close the missing feature decisions

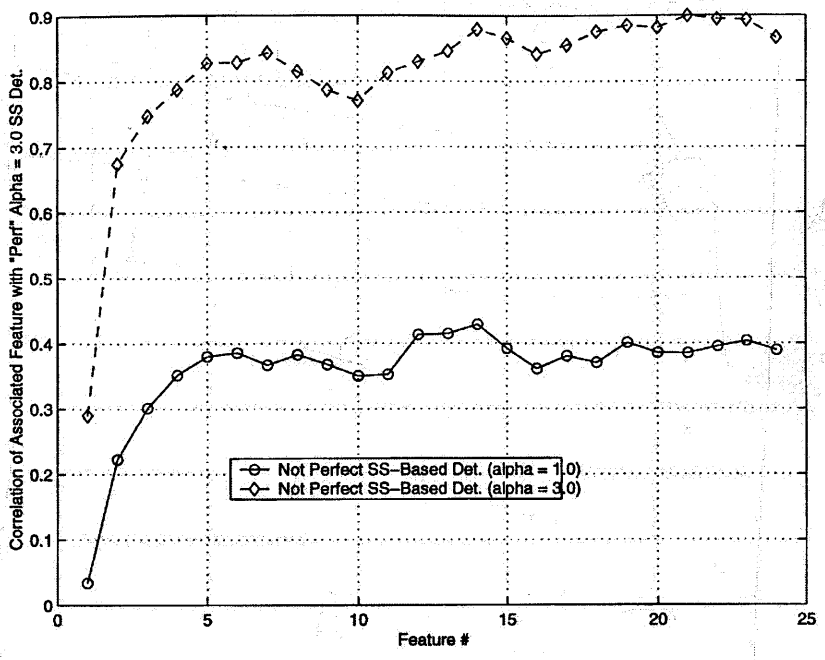

Figure 4-4: Average correlation between the perfect mf detector ($\alpha = 3.0$) and nonperfect mf detectors ($\alpha = 1.0$ and 3.0). Shows that the average correlation for the $\alpha = 3.0$ nonperfect detector is > 0.7 for most of the features, close to 0.9 for the highest ten features. The nonperfect detector with $\alpha = 1.0$ has been shown for reference.

made by the nonperfect detector, using only an estimate of $\overline{\mathcal{N}}_{lin}[m, l]$, are to those made by the perfect detector, which is fortunate to be able to use the actual instantaneous noise component $\mathcal{N}_{lin}[m, l]$. To address this issue the correlation between the perfect and nonperfect missing feature detectors for each feature was studied in an AWGN environment at SNR levels from 10 dB to 20 dB, the average correlations of which are shown in figure 4-4. For comparison, the correlation results for the $\alpha = 1.0$ nonperfect mf detector has been shown as well. This figure demonstrates that for the majority of features (roughly features number four and greater) the correlation between the perfect and nonperfect ($\alpha = 3.0$) mf detectors is rather high, greater than 0.7 and closer for 0.9 for the majority of features. The result of this is that it may be somewhat reasonably expected that at various SNR levels the nonperfect estimator's decisions regarding present and missing features will match those of the perfect mf detector rather closely for most features.

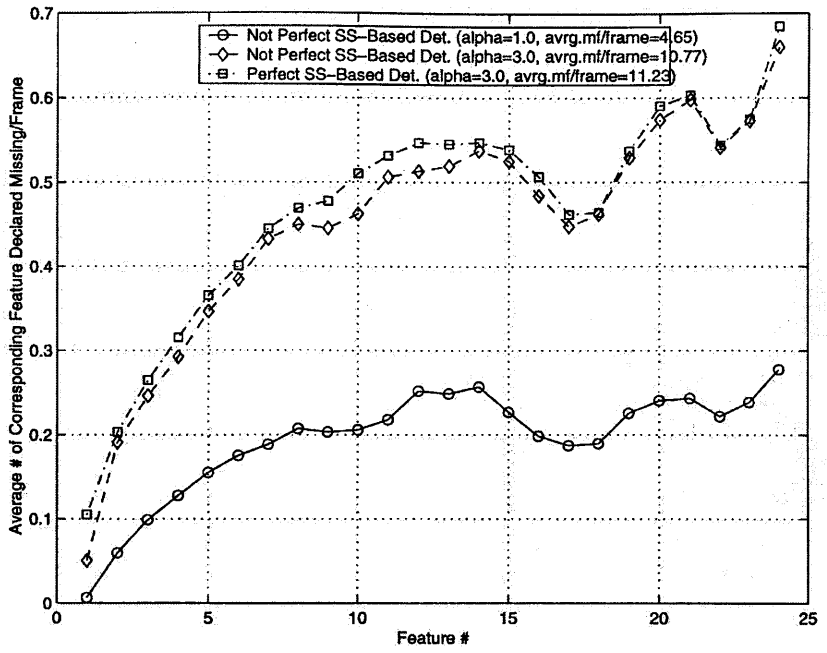

Figure 4-5: Plot showing the average number a particular feature is declared missing per frame by the perfect mf detector($\alpha = 3.0$) and nonperfect mf detectors ($\alpha = 1.0$ and 3.0). The perfect and $\alpha = 3.0$ nonperefect detectors are seen to have produced almost identical results. Results are for 20 dB AWGN.

It is also of interest to know how well the perfect and nonperfect missing features detectors correlate with respect to other aspects of missing features detection, such as the average number of times a particular feature is declared missing per frame as well as the total number of missing features declared missing per frame. These were also measured experimentally for the perfect mf detector and the two nonperfect mf detectors at various SNR values in an AWGN environment. The results showed again that the correlation of the $\alpha = 3.0$ detector with the perfect detector is quite high. The results for an additive noise level of 20 dB may be seen in figures 4-5 and 4-6.

Using these two MF estimators, experiments were run to evaluate the performance of both linear mel-filter energy and $|DFT|$ domain spectral subtraction as well as MFC in comparison to the baseline case. All tests are clean/dirty, where as has been the noise scenario used during this paper the corruption was AWGN. Because in all of these tests it was found that relative performance at all miss and false acceptance

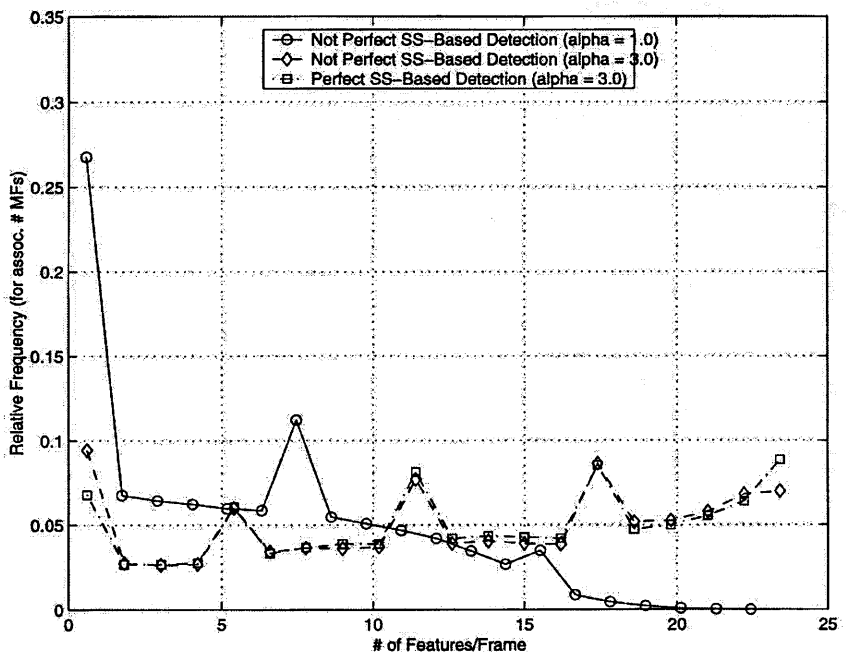

Figure 4-6: Plot showing the frequency per frame that a given total number of missing features are detected by the perfect mf detector($\alpha = 3.0$) and nonperfect mf detectors ($\alpha = 1.0$ and 3.0). The results for the perfect and $\alpha = 3.0$ nonperefect detectors are seen to be almost identical. Results are for 20 dB AWGN.

probability levels was reflected in the relative EER differences, for clarity results will be reported in terms of EER[5].

Results comparing the performance of MFC, with missing features detected with both the perfect and nonperfect estimators, with the baseline case is given in figure 4-7. One important thing to notice regarding the two curves for the MFC case is that they are nearly overlapping, with the maximum deviation being less than 1% at lower SNR values. It is seen that while the application of MFC at lower SNR values tends to hurt EER performance somewhat, at higher SNRs of roughly 17 dB or more the MFC begins to outperform the baseline case. At noise levels greater than roughly 18 dB the improvement over baseline rises to approximately 4%. The reason for a lack of uniform improvement at different SNRs using MFC is likely to be due to the noise vs. speech information discussed earlier in this chapter and highlighted in figure 4-3. In this study it was noted that starting from the case where no corruption is added, as noise is increasingly added to features MFC tends to outperform the case in which no processing is done up to a certain level of corruption, beyond which the application of MFC will begin to hurt performance more than it helps. The explanation for this is that at high SNR levels the amount of speech information retained in the system after MFC is still high enough to effectively do speaker verification; there is an amount of "redundancy" of speech information and the fact that removal of poor features removes some of the noise in the scoring mechanism helps to improve performance. At a point however, the amount of corruption has become high such that more features are being declared as missing and more speech information is being thrown away. At this point the loss of needed speech information degrades performance beyond any improvement resulting from noise removal.

[5]Although it has been observed that relative performance differences between two systems as reflected in EER often reflects relative performance at all the different possible threshold levels described by the DET curve, this is not always the case and care must be taken to not make overgeneral conclusions based on EER.

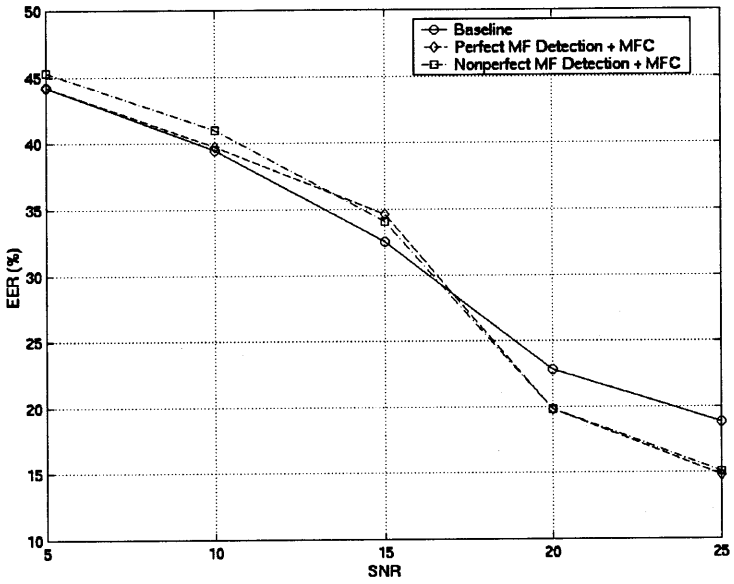

Figure 4-7: EER vs. SNR for MFC with both perfect and nonperfect applications of MFC as well as the baseline case. Both the nonperfect and perfect missing feature detectors are shown to have nearly identical results at all SNR levels. At lower SNRs MFC is seen to degrade performance slightly ($< 1\%$), while at higher SNR levels above approximately 17 dB it starts to improve performance by close to 3.5%.

Chapter 5

Cascade Noise Handling Systems

In the earlier chapters of this thesis various methods of channel compensation and missing feature concepts were presented. The techniques examined have been cepstral mean subtraction, RASTA, spectral subtraction (both linear mel-filter energy and *DFT* domain varieties), and missing feature compensation. For each approach, the effect of the technique was investigated individually.

The natural question, given the earlier results, is what additional performance gains are available if these techniques are combined. Each of these methods is designed to address a particular channel compensation or noise handling need, the underlying assumptions and techniques being different and non-overlapping. In this chapter various combinations of these four building block techniques is investigated. The ability of the separate techniques in each combination to complement each other, paying attention to the noise reduction vs. speech distortion trade-off, determines how well they do when combined. Each system is used in a clean/dirty speaker verification task in which the dirty data was clean speech corrupted with levels of AWGN varying from 5 to 20 dB. In addition, in all cases the selected feature type is logarithmic mel-filter energies and the sampling rate is 8 khz, resulting in 24 features. All GMM speaker models are of order 1024.

5.1 RASTA with Spectral Subtraction

In this section the performance of a system combining RASTA feature trajectory filtering with both linear mel-filter energy and DFT domain forms of spectral subtraction is considered. The results for these two systems, as well as plots for the baseline and constituent systems individually, may be seen in figure 5-1. It is seen that the system combining RASTA with the $|DFT|$ domain spectral subtraction does the best of the two combination systems considered. Compared to the pure $|DFT|$ domain SS system, it only does slightly better ($\approx 2\%$) at low SNR values. As the noise level increases above 15 dB, the system equals and then begins to underperform the pure $|DFT|$ domain SS system. One possible explanation for this is that at low SNR levels the more narrow bandpass RASTA filter may do a better job removing high power noise, necessary to improve performance, while at high SNR values this may eliminate spectral speaker information than noise. Hence the observed performance could be a result of the noise removal vs. spectral speaker information retention trade-off discussed earlier.

5.2 RASTA with Missing Feature Compensation

This section investigates at the performance when RASTA is combined with missing feature compensation. In order to do the MFC, both the perfect and nonperfect missing feature detectors are used, with nearly identical performance. The results are shown in figure 5-2, where it is seen that combining RASTA with MFC either almost equals the performance of pure RASTA (at SNRs greater than 20 dB), or does worse by close to 4% (at all other SNRs). At lower SNR levels, the combination system is seen to actually underperform the baseline case.

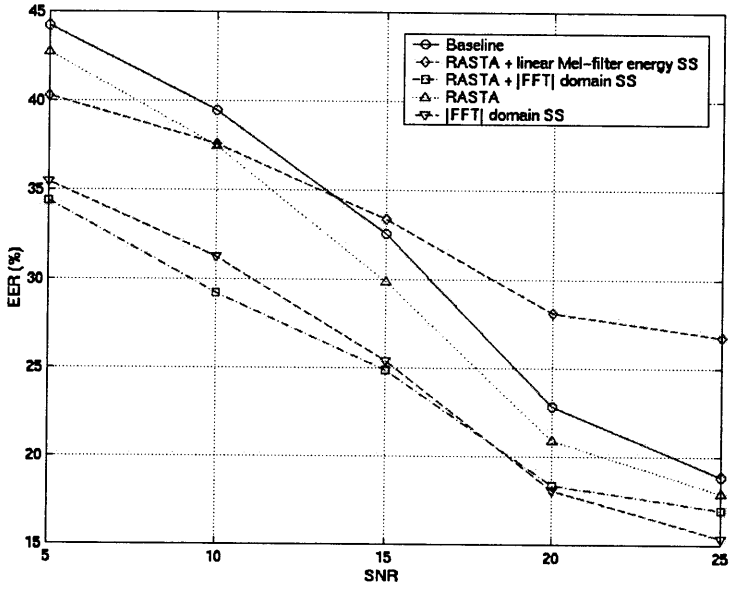

Figure 5-1: EER performance as a function of SNR for cascade systems using RASTA and one of linear mel-filter energy or $|DFT|$ SS systems. For comparison the baseline system, the RASTA system, and the $|DFT|$ system are plotted as well. It is seen that when linear mel-filter energy SS is used with RASTA that performance is worse than straight RASTA at all SNR levels except than the lowest, at 5 dB. This is in contrast to the system using RASTA with $|DFT|$ SS, which is seen to do better than either RASTA or $|DFT|$ SS used alone at almost all SNRs. For this system EER performance is typically 8% better than the baseline system.

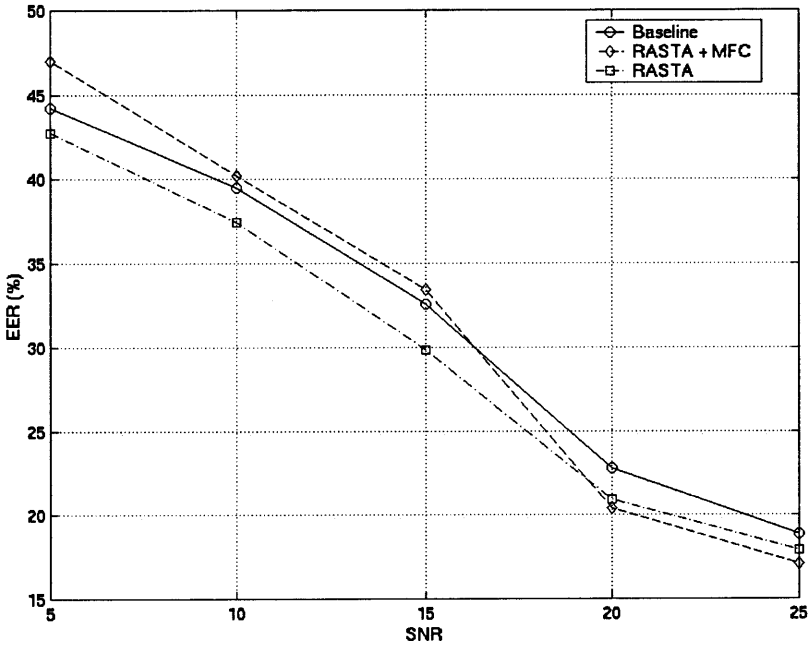

Figure 5-2: Performance with a system combining RASTA with missing feature compensation. Performance curves for the baseline and pure RASTA systems are shown for comparison. It is seen that the RASTA+MFC system underperforms both the baseline and pure RASTA systems, particularly the pure RASTA system. Only at high SNR levels of 20 dB and above are any benefits seen, but then only by about 1%.

5.3 $|DFT|$ Domain Spectral Subtraction with Missing Feature Compensation

Earlier it was seen in section 3.3 that the pure SS system in the $|DFT|$ domain outperformed the performance of SS when applied on the linear mel-filter energies. This section investigates how the performance of these systems changes when they are combined with missing feature compensation. The results are seen in figure 5-3. It is seen that the combination of $|DFT|$ domain SS and MFC greatly outperforms the pure SS systems, the linear mel-filter energy SS and MFC combination system, as well as the baseline. The observed performance improvement holds at all SNR levels and is able to achieve a reduction in EER by an average of as much as approximately 15% at the various SNR levels tested. The second best performing system in this set of trials was the pure $|DFT|$ domain SS system itself, which was outperformed by the combination system at all SNR levels by about 10%.

Of all the various techniques described in this thesis, for the particular channel and training/testing conditions assumed, the system combining $|DFT|$ domain SS and MFC has been shown to perform the best. The resulting performance gain is equivalent to a significant 13 dB gain in improvement as measured by equivalent noise power. Note that it was found that performance was nearly identical for both perfect and nonperfect missing feature detectors.

5.4 RASTA with Spectral Subtraction and Missing Feature Compensation

Finally, this section investigates combining RASTA with spectral subtraction and missing feature compensation. Both $|DFT|$ domain and linear Mel-filter energy domain SS are employed, with results seen in figure 5-4. It is observed that, as has been a trend throughout these results, the $|DFT|$ version of SS significantly outperforms that of the linear mel-filter energies. For the particular channel being considered

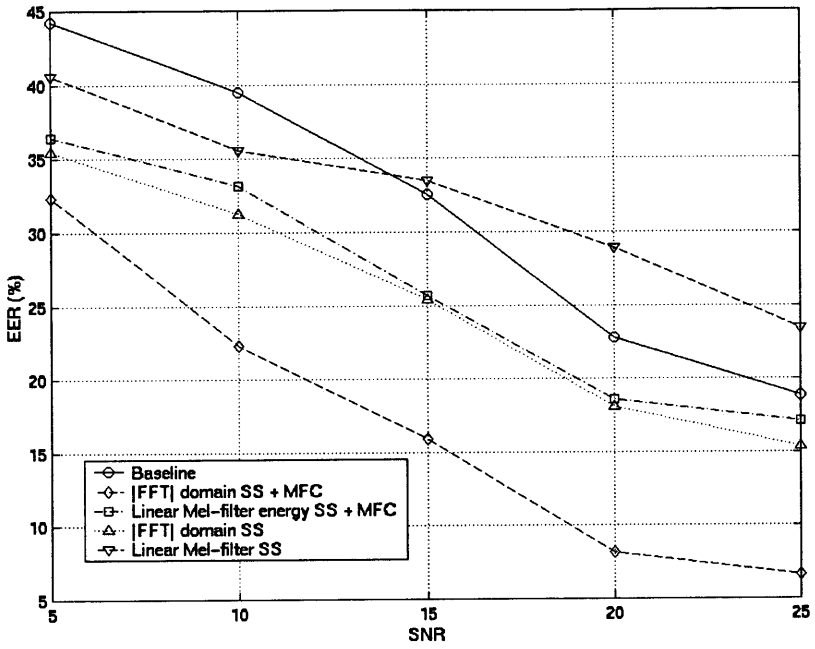

Figure 5-3: Combination systems of $|DFT|$ domain SS + missing feature compensation and linear mel-filter energy SS + missing feature compensation. Also shown are the baseline and pure SS systems' performance for comparison. The $|DFT|$ domain SS + MFC system substantially outperforms the linear mel-filter energy SS + MFC system as well as the other systems shown, at all SNR values. It is seen to perform better than the baseline system by approximately 15%. The linear mel-filter energy SS + MFC system, on the other hand, only achieves performance roughly halfway between these two.

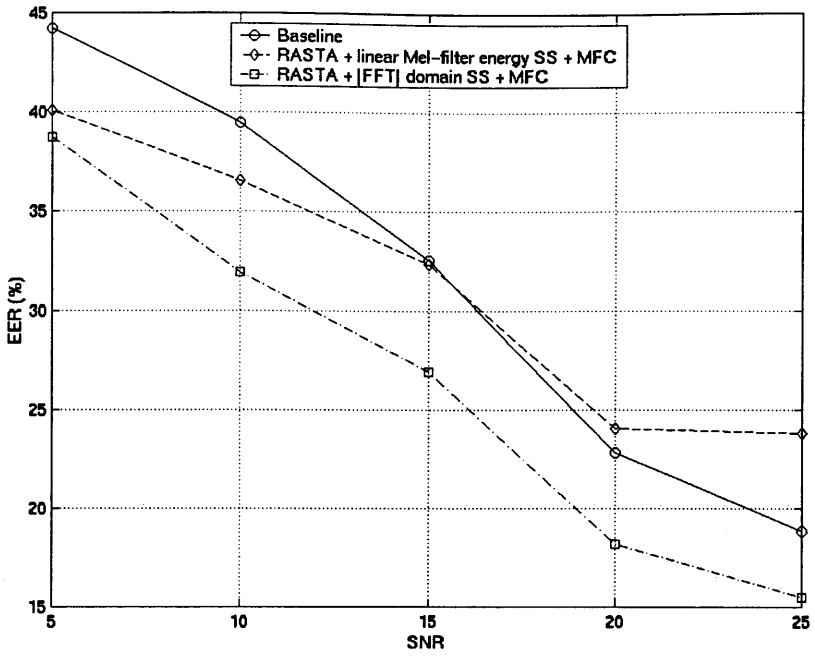

Figure 5-4: RASTA processing in combination with missing feature compensation and linear mel-filter energy SS or $|DFT|$ domain SS. The system with the $|DFT|$ domain SS is seen to do better than the other system as well as baseline for at all noise levels. The system with linear mel-filter energy SS only improves over the baseline at low SNR values, hurting performance at the higher ones.

however (AWGN), the system combining $|DFT|$ domain SS with MFC is seen to do best without additional RASTA processing. For this channel it may be assumed that RASTA processing produced more speech distortion than useful noise reduction. This is in accordance with the intended purpose of RASTA filtering, which is to remove convolutional distortion, not additive noise.

Chapter 6

Missing Feature Restoration

In the previous section the theory of missing features and the technique of missing feature compensation were introduced. In missing feature compensation, the approach in dealing with speech features declared missing is to adapt the automatic speaker recognition system around that void in available speech information to ensure that it does not get included in the scoring mechanism. It is assumed that any speaker information present in the feature is so heavily masked by the noise/distortion that its usefulness or recoverability is negligible and hence should simply be eliminated from consideration in the scoring stage. Missing feature restoration takes a different approach in this scenario by attempting to take advantage of various auto and cross-correlation properties inherent in the different speech feature trajectories and use them to estimate the value of the missing feature. Once the feature's values has been estimated, the estimate may be used in place of the corrupted missing feature.

This concept of missing feature restoration is a new concept in robust speaker recognition and as such is largely unexplored. To date, only one study in this[2] has appeared in the literature. This chapter briefly qualitatively describes the approach and results thus far and then describes the missing feature estimation method that is researched in this thesis.

6.1 Previous Work in MFR

In their paper on missing feature estimation[2], El-Maliki and Drygajlo propose two different ways in which missing features are estimated. This section briefly outlines these two methods and their corresponding results.

6.1.1 Integrated Speech-Background Model

This first method begins with the assumption that the corrupting noise process vector $\mathbf{N}_t$ degrades the underlying clean speech feature vector $\mathbf{X}_t$ in a component-wise fashion producing a noisy feature vector $\mathbf{Y}_t$ through the function f, i.e.,

$$\mathbf{Y}_t = f(\mathbf{X}_t, \mathbf{N}_t).$$

in the linear domain. When transformed to the logarithmic domain, denoted by the l superscript, the MAX model assumed states that the observed feature vector $\mathbf{Y}_t^l$ is equal to the greatest of the logarithms of $\mathbf{X}_t$ and $\mathbf{N}_t$, $\mathbf{X}_t^l$ and $\mathbf{N}_t^l$ respectively, such that

$$\mathbf{Y}_t^l = \log(\mathbf{X}_t + \mathbf{N}_t) \approx \max(\mathbf{X}_t^l, \mathbf{N}_t^l)$$

in which case $f \rightarrow \max$. It is also assumed that the ambient noise may be modeled by a GMM in the same manner as the speech, given by

$$p(\mathbf{X}|\lambda_n) = \sum_{i=1}^{D_{\lambda_n}} p_i^n \Phi_i^n(\mathbf{X}, \mu_i^n, \Sigma_i^n).$$

Given this model, the *integrated speech-background model (ISBM)* missing feature estimator takes the following form

$$\hat{\mathbf{X}}_t^l = E(\mathbf{X}_t^l|\mathbf{Y}_t^l, i, \lambda) = \int\int_C \mathbf{X}_t^l p(\mathbf{X}_t^l|\mathbf{Y}_t^l, i, \lambda_n) d\mathbf{X}_t^l dC_n^l$$

where C is the contour of integration defined by the range of $\mathbf{X}_t^l$ as determined by the MAX model as well as the assumed underlying noise model λ_n. Note that this

estimator is not equivalent to the optimal minimum mean-square error estimator.

6.1.2 Mean Estimation

As mentioned in the section on missing feature theory and compensation, after missing features have been detected the speech feature vector $\mathbf{X}$ may be thought of as being composed of two sub-vectors, denoted by $\mathbf{X}_{present}$ and $\mathbf{X}_{missing}$, which represent the present and missing features. With regards to the speaker model parameters, a similar grouping may be done as follows

$$\mu = (\mu_{present}\ \mu_{missing})^T\ ,\ \Sigma = \begin{pmatrix} \Sigma_{pp} & \Sigma_{pm} \\ \Sigma_{mp} & \Sigma_{mm} \end{pmatrix}.$$

For the purpose of missing feature estimation, an additional GMM model for each speaker is constructed composed of a single Gaussian density. The first method of *mean estimation (ME)* missing feature estimation simply replaces any missing features $\mathbf{X}_{missing}$ with their corresponding mean values $\mu_{present}$ from the single Gaussian pdf speaker model, i.e.,

$$\hat{\mathbf{X}}_{missing} = \mu_{missing}.$$

The second estimation method performs a conditional mean calculation via linear regression using the parameters of the single Gaussian density model:

$$\begin{aligned} \hat{\mathbf{X}}_{missing} &= E(\hat{\mathbf{X}}_{missing}|\hat{\mathbf{X}}_{present}, \lambda) \\ &= \mu_{missing} + \Sigma_{mp}(\mathbf{X}_{present} - \mu_{present})\Sigma_{pp}^{-1}. \end{aligned}$$

It is important to note that this estimator is also not the optimal minimum mean-square estimator.

6.1.3 Results

In [2], with speech taken from the NTIMIT database, a collection of clean phone quality speech, was used to evaluate the above techniques. 400 speakers were selected and for each speaker two sentences were used. The speaker models were 32-order GMMs and the noise added was AWGN.

At SNRs ranging from 0 to 18 dB, both the Integrated Speech-Background Model and Mean Estimation estimation methods were shown to perform worse than Missing Feature Compensation while much better than the case where no missing feature related processing is applied. Compared to MFC, at all SNRs the ISBM estimator resulted in EER values about 2% higher while the conditional mean and mean substitution variants of ME estimation resulted in EER values respectively about 1% and 5% higher. For comparison, MFC alone had EER values typically 15% lower than the baseline case.

6.2 Time-Frequency Linear Minimum Mean-Squared Error Missing Feature Estimation

In considering the ISBM and ME methods of missing feature estimation, it is important to note which values are being used in making the estimates. If a given frame has been detected to contain a missing feature, the two methods above attempt to estimate its value using present features from the same frame. While this is useful in that it makes avail of potential correlations and dependencies in the mel-filter energies across the frequency axis, they completely neglect any sort of time-domain correlations that may exist along each feature's time-trajectory as well as correlations across different features' time-trajectories that may exist. In this section a new approach to missing feature estimation, *time-frequency linear minimum mean-squared error (MMSE) missing feature estimation* is developed. It is most strongly differentiated from the two methods above by its use of present features in both the time and frequency domains, as well as its use of a linearly constrained estimator.

To develop the motivation for the approach being developed in this thesis, it is useful to begin with a short derivation of the optimal mean-square estimator. To set up the problem, consider x_{miss} and $\hat{x}_{miss}$ to be the underlying true value of a speech feature declared to be missing and that feature's estimate, respectively. Let $\mathbf{R} = [r_1, r_2, ..., r_L]$ be the inputs, whichever features they may be, to be used by the estimator. Assume that $\mathbf{R}$ and x_{miss} are both zero mean. Defining the estimation error by

$$\varepsilon = x_{miss} - \hat{x}_{miss}$$

and the cost function of our estimator as

$$\begin{aligned} C(x_{miss}, \hat{x}_{miss}(\mathbf{R})) &= \varepsilon^2 \\ &= (x_{miss} - \hat{x}_{miss}(\mathbf{R}))^2, \end{aligned}$$

the *expected risk* in the estimate may be defined as

$$E[\mathcal{R}] = E[C(x_{miss}, \hat{x}_{miss})].$$

Expanding out $E[\mathcal{R}]$ gives

$$\begin{aligned} E[\mathcal{R}] &= E[C(x_{miss}, \hat{x}_{miss}(\mathbf{R}))] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(x_{miss}, \hat{x}_{miss}(\mathbf{R})) p_{x_{miss},\mathbf{R}}(x_{miss}, \mathbf{R}) dx_{miss} d\mathbf{R} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_{miss} - \hat{x}_{miss}(\mathbf{R}))^2 p_{x_{miss},\mathbf{R}}(x_{miss}, \mathbf{R}) dx_{miss} d\mathbf{R} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_{miss} - \hat{x}_{miss}(\mathbf{R}))^2 p_{\mathbf{R}}(\mathbf{R}) p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R}) dx_{miss} d\mathbf{R} \\ &= \int_{-\infty}^{\infty} p_{\mathbf{R}}(\mathbf{R}) d\mathbf{R} \int_{-\infty}^{\infty} (x_{miss} - \hat{x}_{miss}(\mathbf{R}))^2 p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R}) dx_{miss}. \end{aligned}$$

The estimator $\hat{x}_{miss}(\mathbf{R})$ will be chosen so as to minimize $\mathcal{R}$. In other words

$$\hat{x}_{miss}(\mathbf{R}) = \min_{\hat{x}_{miss}} E[C(x_{miss}, \hat{x}_{miss}(\mathbf{R}))].$$

Because the terms in the first integral above are non-negative for each value of $\mathbf{R}$, the expected risk $\mathcal{R}$ may be minimized by minimizing the second integral. Taking the partial derivative of the second integral with respect to $\hat{x}_{miss}$ and setting expected risk $\mathcal{R}$ equal to 0 gives

$$\begin{aligned}
\frac{\partial E[\mathcal{R}]}{\partial \hat{x}_{miss}} &= \frac{\partial}{\partial \hat{x}_{miss}} \int_{-\infty}^{\infty} (x_{miss}^2 - 2\hat{x}_{miss}x_{miss} + \hat{x}_{miss}^2) p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R}) dx_{miss} \\
&= -2\int_{-\infty}^{\infty} x_{miss} p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R}) dx_{miss} + 2\hat{x}_{miss} \int_{-\infty}^{\infty} p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R}) dx_{miss} \\
&= -2\int_{-\infty}^{\infty} x_{miss} p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R}) dx_{miss} + 2\hat{x}_{miss} \\
&= 0
\end{aligned}$$

and finally rearranging this produces our optimal estimate for the missing feature x_{miss}

$$\begin{aligned}
\hat{x}_{miss} &= \int_{-\infty}^{\infty} x_{miss} p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R}) dx_{miss} \\
&= E[x_{miss}|\mathbf{R}],
\end{aligned}$$

which is the well-known *minimum mean-square error (MMSE) estimator*[14], in this case stated in terms of estimating a missing feature x_{miss} from some number of inputs $\mathbf{R}$.

While being able to calculate this estimate is the ideal scenario due to its optimality for the adopted cost function, it is often going to be impossible (or at least extremely difficult) to do so. In order to calculate $E[x_{miss}|\mathbf{R}]$ it is required that $p_{x_{miss}|\mathbf{R}}(x_{miss}|\mathbf{R})$ be reliably known, which is very difficult to do in that it requires knowledge of all moments of the elements of $\mathbf{R}$ and their joint statistics with x_{miss}.

In most cases getting reliable estimates for these values will be hard to achieve.

In order to limit the estimator to one that requires estimation of only first and second moments and hence may be more realistic, as well as computationally simple, to implement, the estimator will be limited to one that is a linear combination of the inputs. The estimator now has the following form

$$\hat{x}_{miss} = \sum_{k=0}^{L} w_k r_k = \mathbf{R}^T \mathbf{W}.$$

Using the same MSE cost criterion $\mathcal{R} = E(\varepsilon^2) = E[C(x_{miss}, \hat{x}_{miss}(\mathbf{R}))]$ as before, as well as the assumption of a zero mean $\mathbf{R}$ and x_{miss}, the calculation of the optimal estimator reduces to finding the optimal set of weights $\mathbf{W}_{opt}$

$$\begin{aligned} \mathbf{W}_{opt} &= \min_{\mathbf{W}} E[(x_{miss} - \hat{x}_{miss}(\mathbf{R}))^2] \\ &= \min_{\mathbf{W}} E[(x_{miss} - \mathbf{R}^T\mathbf{W})^2]. \end{aligned}$$

Expanding out the cost function C gives

$$\begin{aligned} C(x_{miss}, \hat{x}_{miss}(\mathbf{R})) &= (x_{miss} - \hat{x}_{miss}(\mathbf{R}))^2 \\ &= x_{miss}^2 + \mathbf{W}^T\mathbf{R}\mathbf{R}^T\mathbf{W} - 2x_{miss}\mathbf{R}^T\mathbf{W}. \end{aligned}$$

If we assume that ε, x_{miss}, and $\mathbf{R}$ are WSS[1], then our expected risk takes the following form

$$\begin{aligned} E[\mathcal{R}] &= E[\varepsilon^2] \\ &= E[(x_{miss}^2 + \mathbf{W}^T\mathbf{R}\mathbf{R}^T\mathbf{W} - 2x_{miss}\mathbf{R}^T\mathbf{W})^2] \\ &= E[x_{miss}^2] + \mathbf{W}^T E[\mathbf{R}\mathbf{R}^T]\mathbf{W} - 2E[x_{miss}\mathbf{R}^T]\mathbf{W}. \end{aligned}$$

[1]A constraint that will be worked around shortly via an adaptive linear MMSE estimator.

Redefining the correlation matrices in the above expression as

$$\Psi = E[\mathbf{R}\mathbf{R}^T]$$
$$\Phi = E[x_{miss}\mathbf{R}^T]$$

allows the expected risk to be simplified to

$$E[\mathcal{R}] = E[x_{miss}^2] + \mathbf{W^T}\Psi\mathbf{W} - 2\Phi\mathbf{W}.$$

To find the optimal choice of weights $\mathbf{W}_{opt}$, the partial derivative of $E[\mathcal{R}]$ is taken with respect to the scalar weights $w_1, w_2, ..., w_L$ and set equal to 0

$$\frac{\partial E[\mathcal{R}]}{\partial \mathbf{W}}_{opt} = 2\Psi\mathbf{W}_{opt} - 2\Phi$$
$$= 0$$

and then solved to give an expression for W_{opt}

$$\mathbf{W}_{opt} = \Psi^{-1}\Phi.$$

The optimal linear MMSE estimator is thus given by

$$\hat{x}_{miss} = \mathbf{R}^T\mathbf{W}_{opt},$$

where

$$\mathbf{W}_{opt} = \Psi^{-1}\Phi,$$

assuming that Ψ is non-singular. This equation is often referred to as the *Wiener-Hopf equation*[3].

Earlier an assumption was made that both x_{miss} and $\mathbf{R}$ are zero mean. To accom-

modate the more general case of x_{miss} and $\mathbf{R}$ having nonzero means, the following variables may be defined:

$$\tilde{\mathbf{R}} \triangleq \mathbf{R} - \overline{\mathbf{R}}$$

and

$$\tilde{x}_{miss} \triangleq x_{miss} - \overline{x}_{miss},$$

where the bar notation represents the mean. These variables, defined in terms of $\mathbf{R}$ and $\tilde{x}_{miss}$, as well as an estimate of the mean of x_{miss}, $\overline{x}_{miss}$, allow the final linear MMSE estimator may be given as

$$\hat{x}_{miss} = \hat{\overline{x}}_{miss} + \tilde{\mathbf{R}}^T \mathbf{W}_{opt}.$$

6.3 Proposed System for Missing Feature Restoration

In section 6.2 it was shown that the optimal linear MMSE missing feature estimator had the following form

$$\hat{x}_{miss} = \overline{x}_{miss} + \tilde{\mathbf{R}}^T \mathbf{W}_{opt},$$

where

$$\mathbf{W}_{opt} = \Psi^{-1}\Phi$$

and where now $\Psi \triangleq E[\tilde{\mathbf{R}}\tilde{\mathbf{R}}^T]$, $\Phi \triangleq E[x_{miss}\tilde{\mathbf{R}}^T]$, $\tilde{\mathbf{R}} = [\tilde{r}_1\ \tilde{r}_2\ ...\ \tilde{r}_L]$ is the vector of inputs to the estimator minus their means, and $\tilde{x}_{miss}$ is the missing feature to be estimated and restored, minus its mean. In the missing feature estimators developed by Drygajlo and El-Maliki[2], $\tilde{\mathbf{R}}$ represented only frequency domain features. In this thesis a missing feature estimator is developed in which both frequency as well as temporal features in a time-frequency space around the missing feature are used.

What makes an approach such as this feasible is the existence of correlations between the true values of the feature labeled as missing and the other features around it in its time-frequency neighborhood, on the same trajectory as well as surrounding trajectories. In other words, it depends on the availability of statistical relationships between the speech features in both time as well as frequency.

To investigate the presence of such correlations in the mel-filter energies both the auto and crosscorrelations from various trajectories and their adjacent trajectories for typical clean speech was studied. By observing a typical 3-D plot of speech features as they vary with time and frequency for a typical utterance it can be seen that at certain points in the time-frequency space that the features tend to vary with their neighbors in a manner that suggests correlation between them. Such a 3-D plot is given in figures 6-1 and 6-2 for both linear and log features. A representative linear mel-filter energy feature trajectory is also given in 6-3. It can be seen that in many

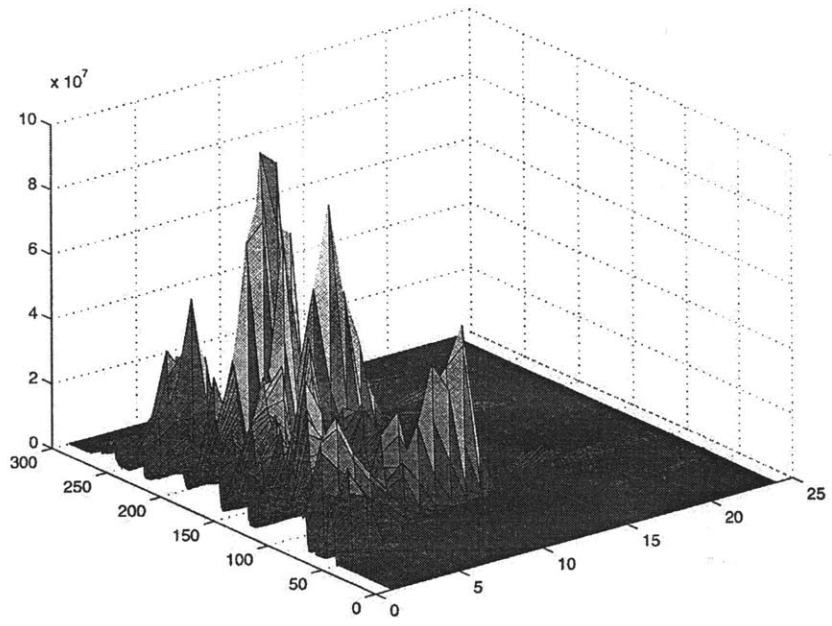

Figure 6-1: 3-Dimensional plot of linear mel-filter energy feature trajectories from a clean speech file.

areas of the plots that if a given point were missing that one might be be able to reasonably extrapolate its value from its surrounding features. It is the hope of MFE to automate this in a mathematically precise way.

Further indicating the correlations in the feature trajectories are several plots which show the autocorrelation functions for four of the mel-filter energy feature trajectories as well as for the crosscorrelation functions between them and their next higher (in frequency) neighbor. These were calculated with clean speech taken from the TSID corpus and are given in figures 6-4 and 6-5.

It is seen in each of the autocorrelation plots that that the features tend to have a hight correlation (> 0.7) within a few frames for each given trajectory. An important point to note is that except for the first mel-filter energy feature trajectory all trajectories display a tendency for a "bump" at a frame lag of typically ± 27 frames or so. Given that the frame rate is 10 ms, 27 lags represent a period of 3.7 Hz, which is very close to the 4 Hz *syllabic rate* that is speculated in the research community to underly most speech. In looking at the crosscorrelation plots it is seen that correlation

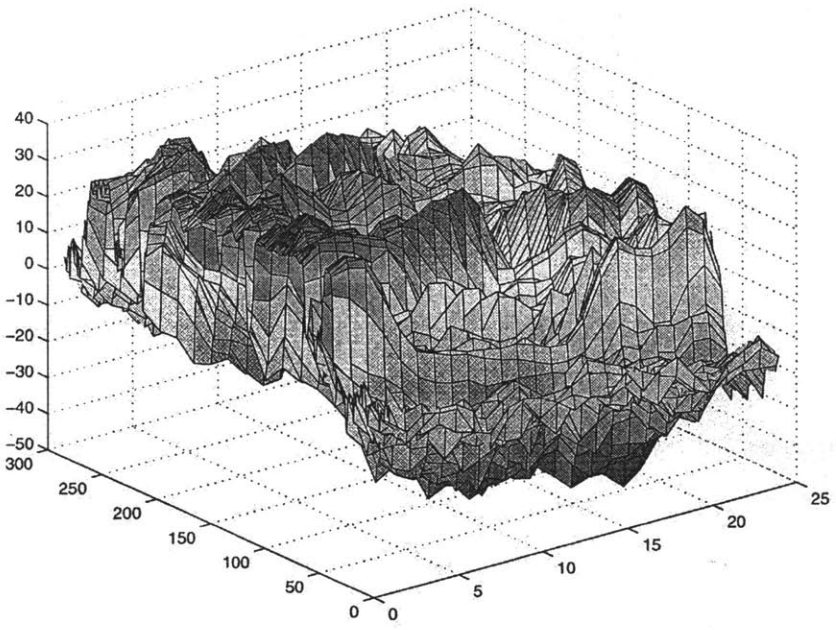

Figure 6-2: 3-Dimensional plot of logarithmic mel-filter energy feature trajectories from a clean speech file.

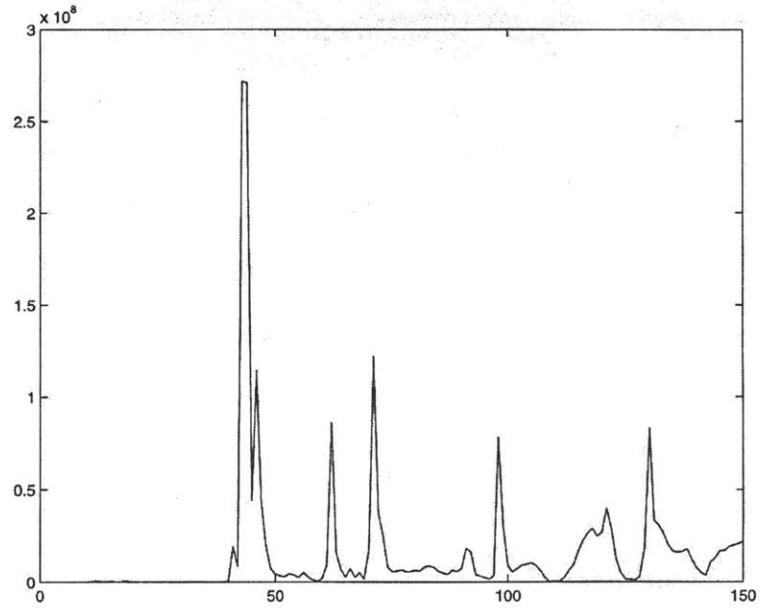

Figure 6-3: Example clean linear mel-filter energy time trajectory. It is seen that the waveform is far from random, having the property that features close in time tend to have similar values and a perceivable underlying deterministic nature.

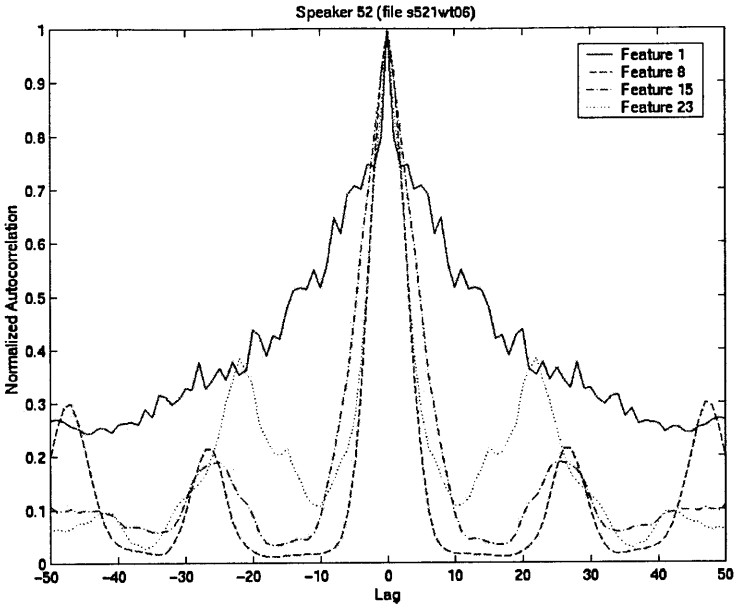

Figure 6-4: Representative autocorrelation function for several linear mel-filter energy features. Speech is taken from a clean speech file.

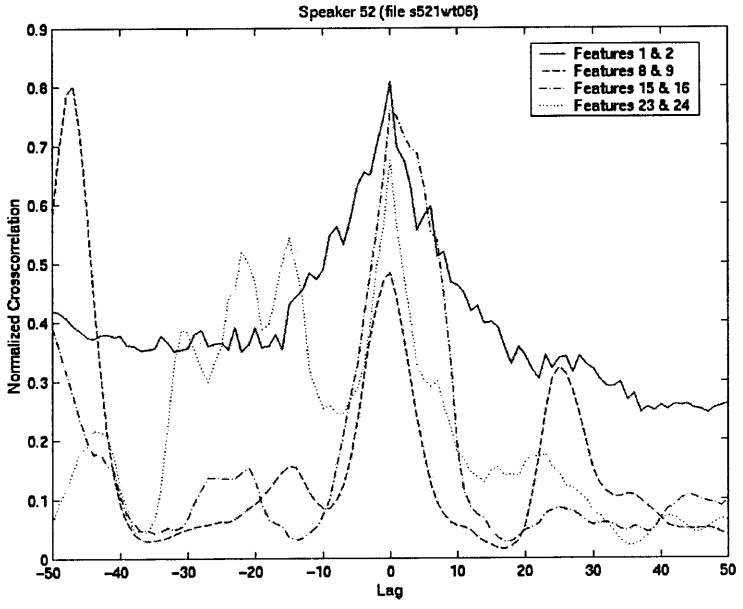

Figure 6-5: Representative crosscorrelation function for several linear mel-filter energy features. Speech is taken from a clean speech file.

of varying amounts also exists across adjacent feature trajectories.

In looking at the auto and crosscorrelation plots it may be seen that the correlation behavior exhibited by the first mel-filter energy is different than the others, in its lack of a syllabic rate "bump" for example. While the true reason for this behavior is unclear, it is likely a result of the bandpass frequency response typical of audio microphones, which was how the clean TSID data was recorded. A small amount of speech energy in the low frequency regions of speech, accompanied by electronic recording noise, would account for the less structured plots for this feature.

While these correlation plots do show correlation that suggests that the proposed missing feature estimator is potentially viable, it raises the question of how the different parts of speech, such as vowels and fricatives, contribute to these overall averages. To explore this question isolated vowel and fricative regions of some typical clean speech files were taken and the corresponding feature trajectories were analyzed for auto and crosscorrelations. The vowel region displayed a high level of auto- and cross-correlation, as may be seen in figures 6-6 and 6-7. In comparison, the correlation plots for the fricative regions demonstrated very little correlation. Those plots are given in figures 6-8 and 6-9.

Given the difference in the nature of voiced and fricative speech regions, it is not surprising that this difference in feature correlation exists. Typically vowels can be thought of as being produced by passing an impulse train with a certain pitch period through a linear system representing a glottal pulse and the vocal tract. Hence vowels may be seen to be periodic and deterministic waveforms with a period equal to the pitch period. Thought of in this way it is almost expected that enough structure exists in voiced regions as to produce the correlations seen above. On the other hand, fricatives are usually modeled by a linear vocal tract model being excited by an AWGN input, resulting in a stochastic colored Gaussian signal. Given this model, the fact that auto and crosscorrelations exist in the corresponding feature trajectories is not outside of reason.

Given that the necessary correlations tend to exist primarily in voiced regions it is clear that the missing feature estimator should be used only for prediction features

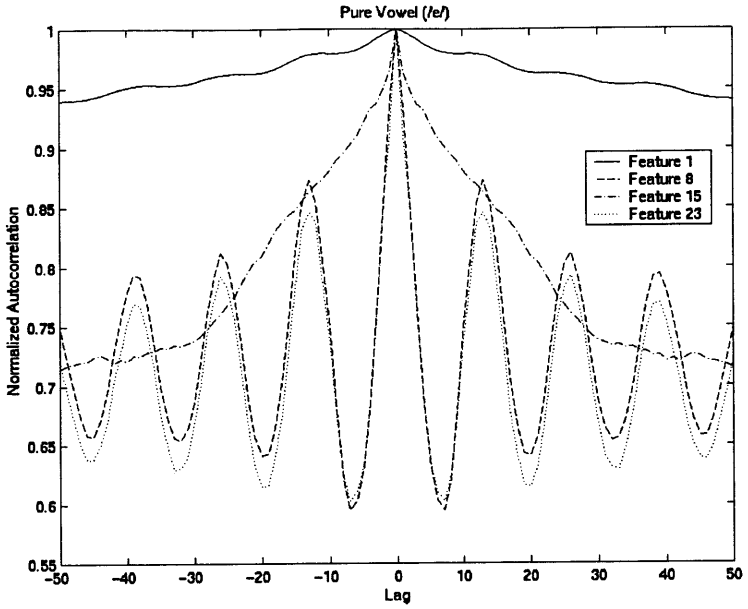

Figure 6-6: Representative autocorrelation function for several linear mel-filter energy features from a pure vowel (/ε/). It is seen that the amount of autocorrelation for each trajectory is rather high, compared to the overall average.

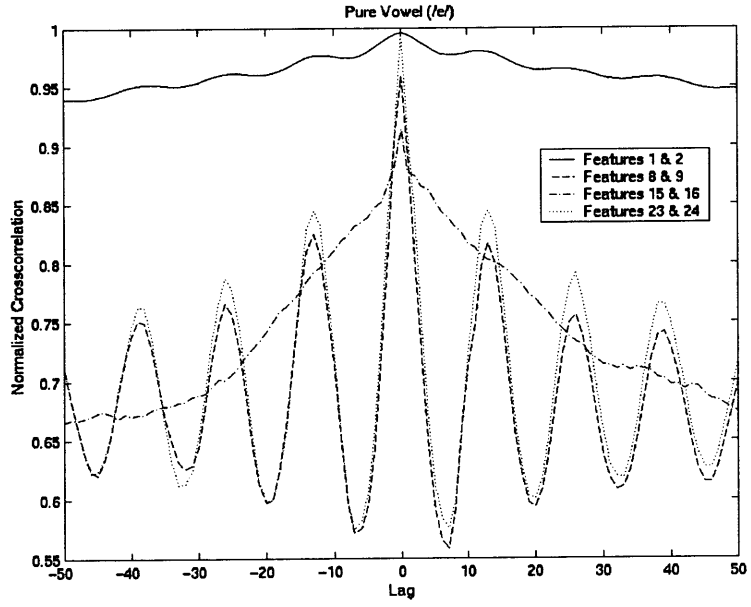

Figure 6-7: Representative crosscorrelation function for several linear mel-filter energy features and their neighbors from a pure vowel (/ε/). It is seen that the amount of cross-correlation for each trajectory is rather high, compared to the overall average.

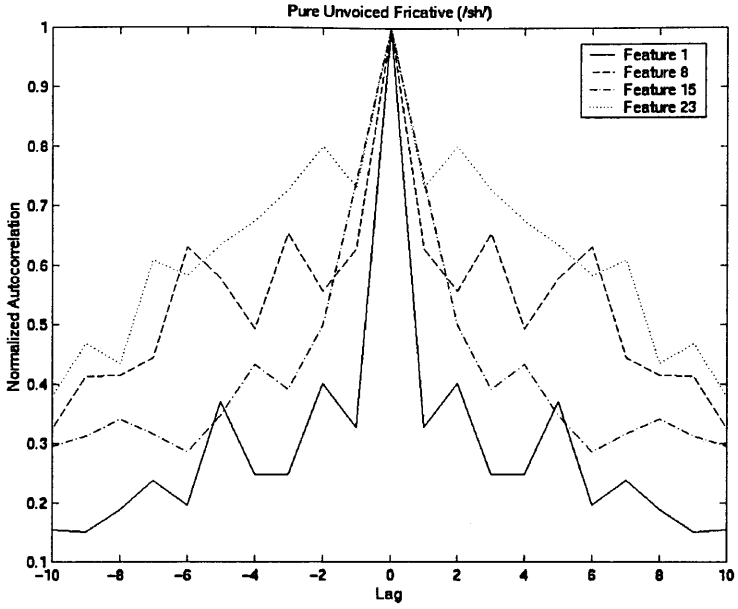

Figure 6-8: Representative autocorrelation function for several linear mel-filter energy features from a pure fricative (/ʃ/). It is seen that the amount of autocorrelation for each trajectory is rather low, compared to the overall average and the pure vowel.

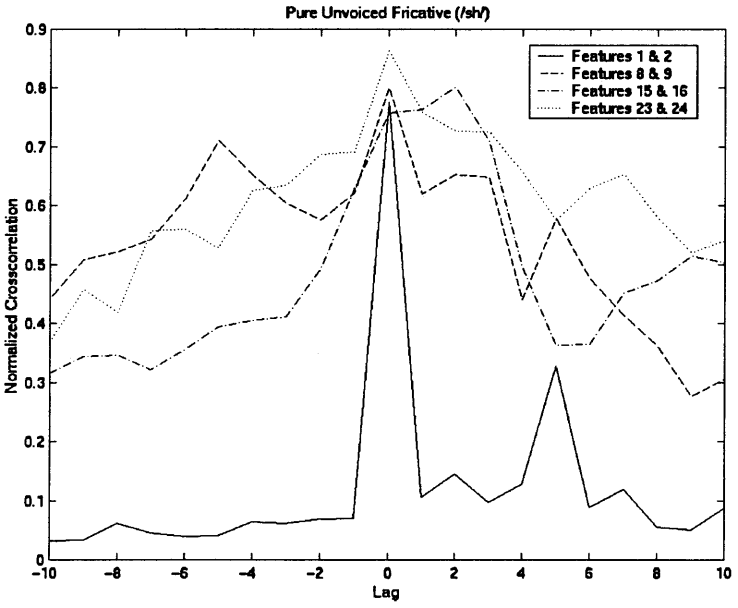

Figure 6-9: Representative crosscorrelation function for several linear mel-filter energy features and their neighbors from a pure fricative (/ʃ/). It is seen that the amount of crosscorrelation for each trajectory is rather low, compared to the overall average and the pure vowel.

in such regions of corrupted speech. In order to do this it will be necessary to detect when in a speech file voiced frames are being encountered and when they are not. To this end a voicing detector is employed which essentially fits a harmonic set of sine waves to the input speech frame using a MSE criterion[7]. The accuracy of the harmonic fit indicates the voicing state and is used to define a "voicing probability" [2] that is between 0 and 1, where 1 indicates a voiced region. If the voicing probability is above a given threshold, the underlying frame is said to be voiced and hence we may use that data for either estimating Ψ and Φ or x_{miss} as the case may be.

Another point to consider is the degree of stationarity of the speech. In the derivation of the optimal linear MMSE estimator an explicit assumption was that x_{miss}, $\tilde{\mathbf{R}}$, and the error ε are all stationary processes. In reality speech is very nonstationary and varies considerably as the speaker weaves in and out of the sounds and transitions between sounds in typical speech. To address this reality of nonstationary speech, the estimation of Φ and Ψ is made to be adaptive such that at any point at time they will reflect as best as possible to most current statistics in the speech features. Initially all voiced frames are pre-processed to develop global estimate Φ_{global} and Ψ_{global}. In subsequent processing Φ and Ψ are initialized to these global values at the beginning of each speech file or when a certain number of unvoiced speech frames have been encountered, done in order to avoid having Φ and Ψ represent outdated statistics, in which case the global estimates would be closer to the current state and a better point to start from when moving into a new voiced region.

Regarding the input vector $\tilde{\mathbf{R}}$, up till now there has been an implicit assumption that all elements of this vector $[r_1\ r_2\ ...\ r_L]$ will always be available, i.e. present. However, just as the feature being estimated could be missing, so could any of the features being used as inputs to the estimator. In order to deal with this issue, when estimating Ψ and Φ in voiced frames the available inputs are used to improve the estimate. However, in the estimation phase whenever missing inputs are encountered

[2]This in not a true "probability" in the sense that 0 indicates no chance of being voiced or that a 1 means that it is sure to be voiced. It is more of a loose term indicating the degree of confidence and is not to be interpreted in terms of more exact probability theory.

the subset of present inputs from $\tilde{\mathbf{R}}$ are used instead to form a new input vector $\tilde{\mathbf{R}}'$ and new correlation matrices Ψ' and Φ' are formed by taking the appropriate elements from Ψ and Φ. In this case the estimator operates as before, but with the primed elements:

$$\hat{x}_{miss} = \overline{x}_{miss} + \tilde{\mathbf{R}}'^{T}\mathbf{W}'_{opt},$$

where

$$\mathbf{W}'_{opt} = \Psi'^{-1}\Phi'.$$

Another issue related to the issue of missing estimator inputs is the question of how many of those inputs are needed in order to make an effective estimate. To address this, a parameter representing the minimum number of inputs needed before the estimator will attempt to estimate and replace a missing feature is incorporated into the system.

Finally, it is important to note that the types of auto- and cross-correlations that exist differ depending on the trajectory in question, as evidenced by the correlation plots seen earlier. For this reason there will be different estimation systems, and thus a different Ψ and Φ, for each feature trajectory.

6.4 Results Using Linear MMSE Missing Feature Restoration

To evaluate the performance of the linear MMSE missing feature detector, two types of studies are performed. All work was done using dirty test speech artificially corrupted with known additive white Gaussian noise. Features used in scoring are logarithmic mel-filter energies and the speech is sampled at 8 khz, resulting in 24 features per frame. All GMM models have 1024 mixtures. The features chosen to be the inputs $\tilde{\mathbf{R}}$ for the estimator are two features in the past and two features in the future in a given missing feature's trajectory as well as the three features symmetric with the missing feature in its two adjacent trajectories were also used. Hence, there is a total number of ten possible inputs.

In the first study, the signal to noise (SNR) ratio of the speech features declared missing and restored was measured both before and after the restoration. Given that both the clean and corrupted linear mel-filter energy features are available, the pre-restoration SNR of all the features both declared to be missing em and able to be predicted may be calculated as

$$SNR_{pre} = \frac{\sum_{t=1}^{T} \sum_{l \in \zeta_{miss}[t]} \mathcal{M}_{true}^2[t,l]}{\sum_{t=1}^{T} \sum_{l \in \zeta_{miss}[t]} [\mathcal{M}_{true}[t,l] - \mathcal{M}_{corr}[t,l]]^2}$$

where $\mathcal{M}_{true}[t,l]$ and $\mathcal{M}_{corr}[t,l]$ are the l^{th} true and corrupted (artificially) linear Mel-filter energy features for frame t, T is the total number of frames, and $\zeta_{miss}[t]$ is the set of features declared by the missing feature detector to be missing in frame t. The corresponding post-restoration SNR, SNR_{post}, is calculated in the same manner with $\mathcal{M}_{corr}[t,l]$ being replaced with the feature value estimated by the linear MMSE missing feature estimator, $\hat{\mathcal{M}}_{true}[t,l]$. It is the goal of linear MMSE missing feature restoration to be able to accurately predict missing features such that $SNR_{post} > SNR_{pre}$.

As discussed in section 6.3, preliminary studies of the linear mel-filter energy feature trajectory correlations suggest that the proposed linear MMSE MF detector

should perform better for voiced vowel regions of speech than fricatives or other types of unvoiced speech. An initial study of the performance in unvoiced regions of speech quickly indicated this to be the case, with SNR_{post} typically being less than SNR_{pre}, indicating that doing nothing at all would have been preferable. To see how well performance can be improved when applied to pure vowel regions, a few recordings of long stationary vowels were artificially corrupted with additive white Gaussian noise and missing feature estimation and restoration was applied. Representative results may be seen in table 6.4.

SNR	MF Det.	MF Pred.	Pred. MF Pre-SNR	Pred. MF Post-SNR
5 dB	88.05%	4.89%	9.36 dB	32.03 dB
10 dB	85.9%	6.4%	9.69 dB	35.04 dB
15 dB	82.0%	13.24%	-3.54 dB	15.17 dB
20 dB	70.0%	10.5%	3.41 dB	18.93 dB
25 dB	63.0%	10.11%	4.49 dB	19.3 dB

Table 6.1: Results of applying proposed linear MMSE missing feature restoration system to a *pure vowel /ɛ/*. Columns indicate additive noise level, percentage of speech features declared to be missing out of all speech features, percentage of features restored out of all speech features, the SNR of all features restored prior to restoration, and the SNR of all features restored after restoration. Results are for an artificially corrupted clean vowel recording.

It is seen in the results that although the particular parameters used in the missing feature detector ($\alpha = 3.0$ and $\beta = 0.0$) resulted in a high percentage, typically close to 80%, of all the speech features being declared missing, due to the voicing and minimum number (4 in these experiments) of estimator inputs available requirements, only 10% of the total number of speech features are actually declared missing, estimated, *and* replaced. For these features, however, the linear MMSE estimator proved to be very effective in replacing the missing features with estimated features that resulted in considerably higher SNR values.

Given these SNR based performance results with pure stationary vowels, the next investigation was how an actual speaker verification task would score using this method. To this end, the same system (same adaptive time constant, etc.) as was

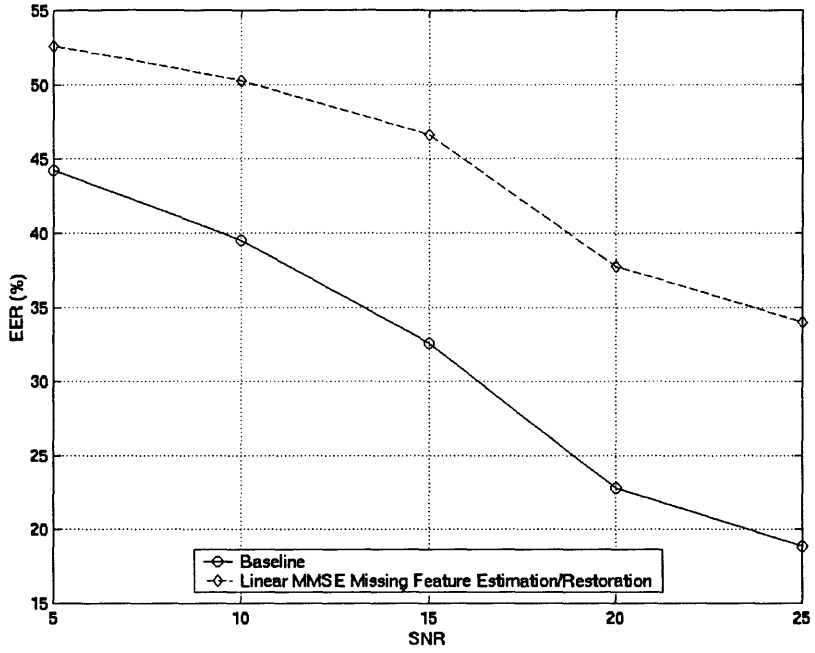

Figure 6-10: EER values for clean/dirty speaker verification task where missing features are detected and replaced, when possible, by using the linear MMSE missing feature estimator. Performance is seen to degrade considerably in comparison to the baseline case.

used to produce the results with the pure vowels was used to also process features in a clean/dirty verification trial, where the dirty test data was a result of corrupting clean speech with artificial Gaussian noise. The results, however, shown in figure 6-10, indicate that when applied to speech, as opposed to a steady voiced vowel, the linear MMSE estimator does not do as well. In fact, it is seen that in most cases performance is degraded significantly.

Hence we see that while the technique does offer a great deal of promise when applied to the pure vowel, in a more realistic scenario of real speech it tends to hurt performance. This is in spite of the fact that the estimation of both Φ and Ψ is being done adaptively and that their estimation as well as the missing feature estimation and replacement is only being done in regions of speech estimated to be voiced. At the time of this writing, the reason for this is still being investigated, but the most likely cause for this discrepancy is that in real speech utterances the voiced regions are possibly too short for the current system to be able to ever acquire adequate

estimates of Φ and Ψ. In the successful pure vowel cases, there was an abundance of data, mostly stationary, by which these correlation matrices could be effectively estimated. With actual speech, however, the vowel regions themselves may not be quite as stationary nor long enough to develop a good estimate. Work needs to be performed to develop better ways of estimating Φ and Ψ in order to realize the potential performance gain that the current results with the pure vowel suggest.

Chapter 7

Conclusions

7.1 Summary of Thesis

The goal of this thesis is to analyze and develop methods for reducing the drastic performance degradation that occurs when the speech used for training and testing in speaker recognition are from different channels. For the TSID speech corpus, in which the "clean" speech was recorded with a microphone near the speaker and the "dirty" speech was a highly corrupted narrowband radio transmission, the train clean/test clean and train dirty/test dirty matched conditions were found to have equal error rate performances of 4.2% and 7.3%, respectively. On the other hand, the mismatched cases of train clean/test dirty and train dirty/test clean both have an equal error rate close to 50%. Given that the decision is binary, this is equivalent to flipping a coin. The purpose of this thesis is to develop methods to bring the performance of the mismatched trials closer to that of the matched trials.

To address this problem, first in chapter 2 several established methods of channel compensation as well as a method for imparting temporal information into the speech features were applied to the train clean/test dirty case[1]. The methods applied were cepstral mean subtraction, RASTA filtering, and delta cepstral coefficients. These methods were applied individually as well as in combination. The best individual

[1]This case is found to perform identically to the train dirty/test clean case.

method was found to be cepstral mean subtraction, which improved the equal error rate from 49% to 28%. The best combination system was the mixture of cepstral mean subtraction and delta cepstral coefficients, which improved the equal error rate to 23%. It was noted that the combination of cepstral mean subtraction and RASTA filtering did the same as the system with only cepstral mean subtraction. One possible interpretation for this is that the largest source of the poor mismatch performance was an invariant convolutional distortion, which either cepstral mean subtraction or RASTA would be able to remove.

Chapter 3 then investigated the theory of spectral subtraction and how this methodology for alleviating the effects of signal corruption, namely additive noise, could be used to help mismatch performance. Because there is a significant amount of additive noise in the dirty TSID data, an ability to reduce the influence of additive noise on the speech features is very beneficial. In order to focus on this type of noise, as well as have control that would allow for a more exact study, clean TSID data was corrupted with known amounts of additive white Gaussian noise to create the dirty data used in testing. Spectral subtraction was applied in both the $|DFT|$ and linear mel-filter energy domains. It was argued that the nature of the nonlinear flooring operation is such that the averaging of this operation offered by the $|DFT|$ domain implementation of spectral subtraction should offer a sort of averaging of this effect, effectively reducing the error component in the values produced by the nonlinearity. Experiments demonstrated that while linear mel-filter energy domain spectral subtraction did offer improved performance over baseline at low SNRs, this was balanced by worse performance at high SNRs. In comparison, the $|DFT|$ domain implementation of spectral subtraction was found to outperform both the linear mel-filter energy version, as well as baseline, at all SNR levels. The equal error rate (EER) was found to be typically 7% lower than for the baseline case.

Chapter 4 discussed missing features theory and the method of missing feature compensation. Missing feature compensation, unlike the previous methods that enhance noise corrupted features, estimates which features are highly corrupted during a given frame and then removes those declared to be missing from being included

in the GMM based scoring mechanism. One concern with this method is that removing the features without doing any proper re-normalization may result in speaker scores varying in a manner uncompensated for by the scoring mechanism, leading to artificially poor performance. It was found that this concern is partially removed by use of background models in normalizing the log probability scores, since both tend to vary together, keeping the normalized score relatively constant. This section also introduced the necessary concept of missing feature detection. While it is not well established that the mean square of the error is the correct value to determine a feature's presence, this metric was assumed to be appropriate throughout this thesis and a spectral subtraction based detector was developed. Through correlation studies it was found that a nonperfect missing feature detector, operating through the estimation of the background noise, produced decisions very close to those of the perfect missing feature detector. It was found that the missing feature compensation system, like the linear mel-filter energy domain spectral subtraction system, produced results that depend on the SNR. At low SNR values it was found that the system underperformed the baseline by about 1%, while at high SNRs it improved the equal error rate by close to 4%.

Chapter 5 next investigated combinations of the previous two chapters, again in a clean/dirty verification task where the dirty speech was clean speech corrupted with AWGN. While it was found that most combinations did better than the baseline, combinations involving $|DFT|$ domain spectral subtraction were found to do considerably well. In particular, the combination of $|DFT|$ domain spectral subtraction and missing feature compensation was found to outperform all other systems considered in this thesis. At all SNR levels it was found to outperform the baseline case with respect to equal error rate by approximately 15%.

Finally, the technique of missing feature restoration was considered in chapter 6. This approach attempts to detect which missing features are missing and then predict their value using other features surrounding it in both frequency and time. Through a study of the auto- and cross-correlations of linear mel-filter energy feature trajectories, it was observed that significant auto- and cross-correlation exists

across features in voiced regions of speech and a linear minimum mean-squared error feature estimator was developed which uses as inputs to the estimator features in the same trajectory, as well as features on neighboring trajectories. Although the pre- and post-SNR results for features that were estimated in this fashion indicated excellent improvements in the case of steady vowels, when applied to regular speech, performance deteriorates dramatically. This occurs even with the use of a voicing detector and adaptive estimation of the correlation matrices used in by the estimator. It is reasoned that a likely source of this error is an inability in the current system to gather enough statistics in each short voiced segment of speech to build reliable correlation matrix estimates. Initial results, however, are promising and more work needs to be done on this problem.

7.2 Suggestions for Future Research

Throughout this research it has been demonstrated that missing feature theory and the related technique of spectral subtraction have value in helping to solve the problem of training and testing mismatch when the the testing data has been corrupted with additive noise. Several future potential advances using these techniques stand out.

Parallel Implementation of Speaker Verification System

In chapter 5 the performance of several noise/channel compensation methods is considered when they are used in combination. In each of the four basic combination systems considered, the constituent systems are acting in series, in which the first system's output becomes the second system's input. This raises the question of what sort of performance might be possible if, rather than operating in series, the constituent systems were instead operated in parallel, with the decisions and/or features from the different systems somehow being combined to make a verification decision. Being in series, it is very possible that systems earlier in the series may process the speech or speech features in such a way as to hurt the ability of later systems in the

series to perform their processing optimally. In other cases, it may instead enhance the processing abilities of later systems. To better understand these relationships parallel system configurations need to be studied.

Preprocessing via Wiener Filtering

Although software compatibility issues made it difficult to study in-depth, one experiment was performance using adaptive Wiener filtering[9] on noisy test data at an SNR of 20 dB for a train clean/test dirty speaker verification task. In this case, the resulting EER was found to be 17.4%, in contrast to the $|DFT|$ spectral subtraction method that had an EER of 18.07% at that SNR. While it is not possible to draw conclusions from a single data point, this experiment as well as well as greater speech quality from Wiener filtering do indicate potential for performance gains equal to or greater than what was seen with the systems involving $|DFT|$ spectral subtraction. In addition, speech data preprocessed with Wiener filtering may smooth out the data such that the performance at later stages of processing, particularly missing feature estimation, may improve.

Missing Feature Estimation in the Log Domain

This thesis implemented the proposed linear MMSE missing feature estimation system on the linear mel-filter energy features. However, it may be shown that the logarithmic mel-filter energies tend to have statistics that are Gaussian-like[5] and, hence, have most statistical relevance in their first and second moments. Given that the linear MMSE estimator only takes advantage of the first and second moments, it may perform better in the logarithmic mel-filter energy domain than in the linear one. In addition, the logarithm's compression may lend itself to smoother feature trajectories and possibly higher correlations. These may lead to improved performance for the missing feature estimator.

Definition of "Missing Feature"

This thesis has defined a missing feature according to the spectral subtraction derived method discussed in section 4.3. However, this is only one way of defining a missing feature, as there are potentially many different and possibly better ways to find which features are most hurting performance. It would be very interesting to investigate the performance differences of the systems described in this thesis with different missing feature detectors to determine the most relevant criterion.

Bibliography

[1] Andrzej Drygajlo and Mounir El-Maliki. Use of generalized spectral subtraction and missing feature compensation for robust speaker verification. In *RLA2C*, April 1998.

[2] Mounir El-Maliki and Andrzej Drygajlo. Missing features detection and handling for robust speaker verification. In *Proc. Sixth European Conf. on Speech Communication and Technology*, volume 2, pages 975–978, Budapest, Hungary, September 1999. The European Speech Communication Association.

[3] Simon Haykin. *Adaptive Filter Theory.* Prentice Hall, 1991.

[4] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[5] N.S. Jayant and Peter Noll. *Digital Coding of Waveforms.* Prentice Hall, 1984.

[6] R.P. Lippman and B.A. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In *Proc. Fifth European Conf. on Speech Communication and Technology*, volume 1, pages KN 37–40, Rhodes, Greece, September 1997. The European Speech Communication Association.

[7] Robert J. McAulay and Thomas F. Quatieri. Pitch estimation and voicing detection based on a sinusoidal speech model. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 249–252, 1990.

[8] Thomas F. Quatieri. *Principles of Discrete-Time Speech Processing.* Prentice Hall, 2000. To be published.

[9] Thomas F. Quatieri and Robert A. Baxter. Noise reduction based on spectral change. In *1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, October 1997.

[10] Douglas A. Reynolds. Automatic speaker recognition using gaussian mixture speaker models. *The Lincoln Laboratory Journal*, 8(2):173–192, 1995.

[11] Douglas A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proc. Fifth European Conf. on Speech Communication and Technology*, volume 2, pages 963–966, Rhodes, Greece, September 1997. The European Speech Communication Association.

[12] Frank K. Soong and Aaron E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(6):871–879, June 1998.

[13] Q. Summerfield, A. Sidwell, and T. Nelson. Auditory enhancement of changes in spectral amplitude. *Journal of the Acoustic Society of America*, 81(3):700–708, March 1975.

[14] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory - Part I.* John Wiley and sons, Inc., 1968.

[15] Sarel van Vuuren and Hynek Hermansky. On the importance of components of the modulation spectrum for speaker verification. In *Proc. ICSLP*, November 1998.